CREDIBILITY FOR CLASSIFICATION RATEMAKING VIA THE HIERARCHICAL NORMAL LINEAR MODEL

STUART KLUGMAN

Abstract

In the past twenty years there has been ever increasing improvement in the techniques of classification ratemaking. Most of this has centered around improvements in credibility procedures and most of the improvements have been due to incorporating aspects of Bayesian analysis. In this paper, I attempt to take this trend to its (perhaps) final stage by developing a true Bayesian approach to the classification ratemaking credibility problem.

The opening section will provide the rationale for the Bayesian approach. I will argue that a hierarchical model with a noninformative prior is the most appropriate general framework. I will argue further that a normal model is a reasonable choice, and this model will provide results at least as good as those currently available. An indication of how the normality condition can be relaxed will also be presented.

The second section contains a general description and analysis of the hierarchical normal linear model (HNLM). Included are point estimation, estimation of the error in the estimator, and prediction intervals for future losses. The last two items are of special interest since current credibility procedures provide little insight with respect to variation.

The next two sections discuss the special case of the one-way model. This is the most common ratemaking model and is the simplest case of the HNLM. In Section 3, the formulas from Section 2 are evaluated for this model. In Section 4, two data sets are analyzed. The first set provides an indication of the computational work required to use the HNLM. The second set provides a comparison of this method with two other ratemaking approaches.

The final section contains a discussion of the more complex models that can be handled with the HNLM.

The majority of this work was supported by a grant from the Society of Actuaries Research Development Fund through the Actuarial Education and Research Fund grants competition. An earlier version of the one-way model was presented at a NATO sponsored conference (Klugman [22]). Many conversations with Glenn Meyers and Gary Venter helped refine the arguments presented here. Two anonymous referees from the CAS provided a number of suggestions that improved the exposition. I thank the National Council on Compensation Insurance and Glenn Meyers for providing the two data sets and especially thank Kathy Hockenberry for writing the programs to do the analyses.

1. JUSTIFICATION FOR BAYESIAN CREDIBILITY WITH A NORMAL MODEL AND A NONINFORMATIVE PRIOR

The historical basis for credibility procedures is long, varied, and generally considered to be one of the major actuarial contributions to statistical data analysis. Virtually from the beginning (Whitney [35] and Bailey [1]), the Bayesian and shrinkage nature of the problem was recognized. In a breakthrough paper, Bühlmann [6] placed the credibility problem in the framework of Bayesian decision analysis. I will begin by reviewing the Bayesian view and then discuss the four schools of Bayesian methodology that are prominent today. As part of this paper, I will argue that one of these methods is superior to the others. Next, I will argue that the normal model is appropriate even though we know that it does not accurately model insurance losses. This part closes with a suggestion for allowing for non-normal losses while retaining the advantages of normal theory. The final element of this section is a discussion of the noninformative prior.

1.1 Credibility as a Bayesian Problem

The basic credibility problem for classification ratemaking can be posed as follows: The population can be separated into k groups, the various rating classes. Our objective is to estimate the mean loss per year generated by a randomly selected member of a particular group. Data is collected from a sample of members from each group. It is usually assumed that the observed losses are independent and that the variances of the observations are proportional to some measure of exposure. If this were all that were known, the most reasonable answer would be to use the sample mean from each group as an estimate of the population mean. Usually, however, we know more. In particular, when individual classes have abnormally good or bad experience, we tend to discount the experience when setting rates. This clearly makes good business sense and with the correct model makes good statistical sense.

The usual way to model this phenomenon is to treat the group means as a random sample from some probability distribution. This implies that experience from the other groups tells us something about the overall level of claims (the mean of this second level distribution), and therefore tells us something about the mean for the group in question. It also sets bounds on how much one can legitimately expect one class to differ from another. If more is known about the relationship among the groups, that knowledge can be incorporated into the second level distibution. Examples of this are presented in Section 5.

The model described in the previous paragraph is a standard Bayesian problem. We have a model given by the p.d.f. $p(x|\theta,g)$ where x represents the data and (θ,g) represents all unknown parameters. The parameters in θ are the ones we want to estimate. The parameters in g are nuisance parameters, usually variances. In the above setting, θ would be the group means. The prior (second level) p.d.f. $p(\theta,g)$ represents our knowledge of (θ,g) before the data are collected. Since the Bayesian approach has now been widely accepted among actuaries (at least for this estimation problem), I will provide no further arguments to support that view. Interested readers who desire a wide ranging discussion of the merits of the Bayesian view are referred to Berger [2].

Given this setup, there are two ways to proceed. If the forms of the two distributions are known, the Bayes estimator is the posterior mean of θ given the data x. Bühlmann [6] took a different approach. To avoid thinking about the distributions, he first restricted himself to estimators that are linear functions of the data. He then searched for the estimator that minimized the mean squared error. This mean would be taken over all possible values of x and θ. For his result it was essential that g be empty. That is, the model variance had to be known. Under this framework, it turned out that the estimator depended only upon the first two moments of the model and prior distributions. To many people, the word credibility is now reserved only for procedures that find linear estimators. In fact, Hewitt [14] compares a credibility estimator to a Bayes estimator (as I have defined it above). In this paper, the objective is to find the best estimator, and I see no reason to restrict attention to those that are linear functions of the data. I use the word credibility to describe any procedure that uses information ("borrows strength") from samples from different, but related, populations.

A larger problem is the fact that the moments of the model and prior are rarely known, and therefore must be estimated. This has led to a number of schools of Bayesian thought. Having agreed to use a Bayesian procedure, the remaining task is to identify the best one.

1.2 Four Schools of Bayesian Analysis

There are at least four different approaches that are currently being used to solve the estimation problem. In this section I briefly outline them and then offer some opinions as to their respective merits.

1.2.1 Pure Bayes with Two Levels

This is the view that has already been mentioned. Here, the prior distribution must be elicited. This is very difficult to do in the insurance setting as one would have to be able to set out a distribution that describes the class-to-class variation in losses. Since we do not even know the means (determining them is the point of the exercise), it is unlikely that we know much about how the means vary.

This problem can be resolved by removing the prior to a higher level of abstraction. This is done in the fourth school discussed in this subsection. To my knowledge, no one today is using the two level approach. At best, it is a starting point for the second method to be discussed here.

1.2.2 Empirical Bayes

This method evolved as an attempt to resolve the problems created by the first method. Although they did not use the phrase "empirical Bayes," Bühlmann and Straub [7] were the first to employ this method in the credibility setting. It remains popular, being advocated in more recent articles by the Insurance Services Office [16] and Meyers [25]. There is considerable evidence that it provides excellent solutions to the estimation problem.

Many people do not consider empirical Bayes methods to be at all Bayesian. Also, there is considerable disagreement as to what the phrase "empirical Bayes" means. To avoid controversy, I will describe an estimation method that corresponds to the approach used in the papers cited above. It will be referred to as the EB approach and the reader can decide what that means. Begin with the density $p(x|\theta,g)$, the first level density (or distribution, when talking about the random variable). The density $p(\theta,g|h)$ will be referred to as the second level density. Note the introduction of h. To the pure Bayesian, the parameters of the second level density must be known, and therefore do not need to be displayed. In reality, that is not true, so we add them to the formulation.

In brief, the EB idea is to first act as if g and h were known and find the Bayes estimate of θ. Next, use the data in some manner to estimate the nuisance parameters g and h and insert these estimates into the Bayes solution. The first thing to note is that upon doing so, we no longer have a Bayesian analysis. The second level distribution was supposed to represent prior opinion, yet here we are unable to establish this distribution until after we have seen the data. The usual justification is to show that as the sample size goes to infinity, the estimates of the second level distribution converge to what they ought to be if we had complete knowledge (which is what one has with an infinite sample size). A thorough discussion of these principles can be found in Norberg [30].

An alternative approach can yield the same solution as the EB approach. Assume again that the nuisance parameters are known and then search for the estimator that is linear in the data and minimizes the expected squared error. Once again, substitute ad hoc estimates of the nuisance parameters into the solution. This has been called least-squares credibility.

There are three major objections to the EB approach. The first is that some external theory must be used to find estimators of the nuisance parameters. Since these parameters are usually variances, it is common to begin with sums of squares that look "right" and then to adjust them to create unbiased estimators of the various parameters. One drawback is that the resulting estimators (even in the simplest cases) can take on negative values. This does not make sense when one is trying to estimate a variance. The second objection is that EB theory gives no guidance as to the optimal choice of the estimator. All that is required is that they be consistent. The final objection is that for complex models, there may be no hope of finding useful sums of squares.

A final problem with EB methodology is that it gives no insight into the sampling error of the estimator. The best it can do is evaluate the error when the variances are known. The additional error introduced by estimating the variances cannot be accounted for. Even if a good estimator of the nuisance parameters can be found (in which case, the method works quite well), the investigator will have no idea of the quality of the estimate. The previous statement that EB methods work well was in reference to alternative methods and does not mean that the results could be considered accurate. That can only be determined by some measure of sampling error. The next method is an attempt to rectify this problem without leaving the EB framework.

1.2.3 Parametric Empirical Bayes

To see the difficulties in determining the variance of the estimator, we need to take a closer look at what we are trying to do. The general Bayes problem is to find $E(\theta|x)$, the posterior mean given only the data. The EB approach uses the result $E(\theta|x) = E[E(\theta|x,g,h)]$. The interior expectation $E(\theta|x,g,h)$ is just the pure Bayes solution with g and h known. The EB approach avoids taking the outer expectation and instead replaces g and h with their estimates. EB theory indicates that this is a reasonable thing to do. A measure of the quality of the result would be the posterior variance, $\mathrm{Var}(\theta|x)$. We have $\mathrm{Var}(\theta|x) = \mathrm{E}[\mathrm{Var}(\theta|x,g,h)] + \mathrm{Var}[\mathrm{E}(\theta|x,g,h)]$. It is apparent that merely inserting estimates of g and h in $\mathrm{Var}(\theta|x,g,h)$ will underestimate the desired variance. The second term reflects the additional variance due to the estimation of g and h. EB theory does not provide any ideas for estimating the second term.

An attempt to resolve this problem is the parametric empirical Bayes theory of Morris [27]. The key ingredient is to have some idea of the variability of the estimators of g and h. His theory requires not only the discovery of good estimators of g and h but also the ability to determine their sampling distributions. In simple cases (normal distribution, equal exposures), it is possible to show that the usual estimators have chi-square distributions. In slightly more complicated cases, the distribution is approximately chi-square. A detailed discussion of the distribution of some commonly used variance estimators is given in Klugman [21]. As should be apparent, there are considerable difficulties associated with putting this method into practice. One that is not apparent, and is often not mentioned in Morris's articles, is that to complete the calculation it is necessary to formulate a prior distribution for g and h. Morris uses $p(g,h) = 1$, but does not provide a justification for that choice. Other choices are supported by an argument that the resulting estimator of the credibility factor is unbiased, a surprising justification for a Bayesian.

The fourth model also requires prior distributions for g and h but proceeds in a more direct Bayesian manner. Before moving on, I should add one final criticism of the parametric empirical Bayes approach. Whatever errors in estimation are introduced cannot be reduced by improving the computational aspects of the method. The errors are due to lack of knowledge of the exact distribution of the estimators of the variances and no amount of computation can resolve that issue.

1.2.4 Hierarchical Bayes

We have seen that the two EB schools are somewhat artificial attempts to resolve the problems of the pure Bayes method. This is mostly due to a lack of recognition of the real problem with the pure Bayes approach. The problem is that the second level distribution is not a prior distribution at all, but is part of the model. In the ratemaking setting, this distribution contains our knowledge of the relationships among the various rating classes. It is not our prior opinion about a particular class. The solution is to reformulate the model into three levels.

Level 1: $p(x|\theta,g)$—Describes variations within each group.
Level 2: $p(\theta|\mu,h)$—Describes variations among the groups.
Level 3: $p(\mu,g,h)$—A true prior distribution on the unknown parameters.

This is once again a pure Bayesian problem. As with any Bayesian analysis, a prior distribution must be established before any data is collected. By displacing the prior to a level further removed from the observations, the choice of the prior will have less influence on the final outcome. To repeat, level 2 describes an underlying (though not directly observable) physical process. Subjective beliefs enter only at the third level. The remaining problems are to select the prior distribution and to select the form of the p.d.f.'s for levels 1 and 2.

Assuming the two problems just mentioned can be resolved, this would appear to be an ideal solution to the credibility problem. With the three densities in hand, it is just a matter of employing the probability calculus to obtain the posterior distribution of θ given x. Any difficulties that will be encountered will be of a computational nature. Once the posterior distribution has been obtained, additional computation will yield the mean and variance. Another useful quantity is the predictive distribution, the p.d.f. (or the mean and variance) of the next observation from the group in question. This is once again obtained by an application of the probability calculus. The other methods do not provide this item.

An additional advantage of this approach is that the tools of Bayesian modeling and inference are all available. For example, one might want to compare various models for the level 2 distribution (e.g., cross-classification vs. one-way classification). Many of these tools are presently in the development stage, but more and more techniques are likely to become available in the future.

1.3 The Normal Model

As mentioned in the previous subsection, it is necessary to specify the probability distributions for the three levels. For levels 1 and 2, multivariate normal models are an appropriate choice. At the end of the section, a suggestion for improving the process is proposed.

It is obvious that individual losses do not follow the normal distribution. Since losses are non-negative quantities, a distribution with support on the entire real line cannot be expected to be a good model. Furthermore, there is considerable evidence that the tails of loss distributions are much heavier than those of the normal distribution. See Hogg and Klugman [15] for a number of examples. One way to minimize the disparity is to work with loss ratios. The distribution at the second level will now reflect group to group variations in the departure from the expected losses. This will be more stable than the group to group variation in the absolute level of losses. In addition, loss ratios are likely to have identical unconditional distributions. That is, if you were given a list of risk classes and a list of loss ratios, you would be unable to do better than chance in attempting to match them up. The loss ratios are likely to be dependent. Knowing that the loss ratio for one class is high increases the chances that the others are also. The multivariate normal model is one of the few multivariate models that allows for dependence in a manner that is easy to construct and interpret.

Despite the fact that the observations are not normally distributed, there are a number of good reasons for employing the normal model. The first, though least appealing, justification is computational convenience. Although the algebra is tedious, as demonstrated in Section 2, a number of results can be obtained analytically. The remaining numerical work will be simple, at least relative to that required for non-normal models. It is likely that as our numerical capabilities increase, this argument for normality will lose its validity. For the present, the following quote due to Novick and Jackson [31] is appropriate.

> "Surely it is better to get some results using a model which is only approximately relevant than to sit twiddling one's thumbs in front of a model which is felt to be more accurate but which one is unable to manipulate."

The second justification for normality is related to the link between the normal model and linear credibility. It was mentioned above that in a particular simple model, the linear least squares solution depended only upon the first two moments. It turns out that the Bayes solution for the same model with normal distributions is identical. Therefore, at least in this case, normality and linear least squares are equivalent. It has been shown that, in general, any model that is a member of the linear exponential family of distributions will produce the same result as the linear least squares solution (Ericson [9] and Jewell [17]). There has been speculation (Goel [11]) that the linear exponential family contains all the distributions with this property. So, to a certain extent, those who are willing to accept linear solutions should be equally comfortable with models from the linear exponential family. As far as choosing the normal distribution as the member to use, a second argument is needed. Most current practitioners estimate the variances using sums of squares. These estimates are unbiased for the variance regardless of the distribution, but any optimality properties relate only to the normal model. The bottom line here is that those who do EB (or least squares) credibility are acting as if they have a normal model without knowing (or admitting) it! By accepting the normal model from the start, we accomplish two goals. First, we know the exact nature of any approximation (normal vs. the true distribution) and second, we have access to all the benefits of the HNLM described in Section 1.2.4.

A third argument for the normal model is that the observations are often not individual losses, but rather the average loss (or loss ratio) experienced by all of the insureds in a particular class in a given year. The central limit theorem indicates that these averages will be more normal in distribution than are the individual observations.

This last item suggests a way to avoid the normal model (to some degree) at the first level. Typically, the level 1 model is that x_{ij} (the j^{th} year average from class i) has a normal distribution with mean θ_i and variance σ^2/P_{ij} where P_{ij} is some measure of exposure. Here the normality comes directly from the central limit theorem. A way to get a better normal approximation would be as follows: Postulate some realistic distribution model (e.g., lognormal or Pareto) for $x_{ij1}, \ldots, x_{ijn_{ij}}$, the individual losses in year j for class i. Let y_{ij} be the maximum likelihood estimator of the mean of the selected distribution (for typical heavy-tailed distributions this will not be the sample mean). Asymptotic theory tells us that y_{ij} has an approximate normal distribution with mean equal to the population mean (θ_i) and a variance that can be estimated from the data. We still have a normal model at level 1, but it is likely to be more accurate than one based on the sample mean. This idea was briefly explored in Klugman [20], but considerable work needs to be done to formalize this procedure and verify its superiority. One problem with this approach is that the estimates are unlikely to be in balance; that is, the total of the estimated means will not match the total observed losses.

1.4 The Noninformative Prior

This is clearly the most difficult part of any Bayesian specification. The prior, or level 3 as it is being called here, must be completely specified. The use of sample information is not appropriate if we hope to recover all of the useful information that a Bayesian analysis can provide. For those who are uncomfortable with the subjective nature of the prior, it should be kept in mind that subjective decisions are made in any statistical modeling procedure, particularly with respect to the choice of the model and the estimation technique to use. In Section 4, an indication of how one might "verify" the choice of model and prior is presented.

Three approaches can be taken to specifying $p(\mu,g,h)$. The first is to always use $p(\mu,g,h) = 1$. Morris [27] uses this prior in obtaining his parametric empirical Bayes results. The first thing to note is that since the support of (μ,g,h) is usually unbounded in at least one direction, this prior is not a proper probability distribution. Box and Tiao [3] argue that this is acceptable. Suppose μ is the average loss ratio over all rating classes. We can be virtually certain that this value is between 0 and 10. A uniform distribution over this interval combined with one that tails off slowly outside this interval would reflect the fact that very little is known about the true average loss ratio. Inferences that we would make using this prior would differ very little from those made using $p(\mu) = 1$ for $-\infty < \mu < \infty$. Two features of this approach should be noted. First, there is no guarantee that the posterior distribution of $\theta|x$ will exist. This would make it impossible to determine the posterior mean or variance. Second, the posterior mode is identical to the maximum likelihood estimator (after integrating out all nuisance parameters). In general, when this prior admits a solution it is quite reasonable.

The second school of thought is to find a general way of obtaining prior distributions that reflect minimal prior knowledge. Words such as "noninformative" or "reference" are often attached to such priors. The goal of research in this area is to find a way to automatically generate *the* noninformative prior for a given distribution. The fact that there is still disagreement on the appropriate reference prior for the probability of success in a sequence of Bernoulli trials (Geisser [10]) indicates how much work remains in this area. In the simple univariate case, Box and Tiao [3] support the prior $p(g) = 1/g$ when g is the variance. An extension is given by Tiao and Zellner [33] who argue that if g is a covariance matrix, the appropriate prior density is the inverse of its determinant.

The third belief is that only proper densities (those that integrate to 1) should be allowed for the prior distribution. Proponents of this approach insist that everyone has a prior distribution and it is just a matter of care and effort to bring it out. This makes excellent theoretical sense but is very difficult to implement. It is even more difficult to convince someone else that your opinion, as expressed by your prior distribution, is valid.

I have elected to take the middle ground. For credibility problems, the reciprocal prior for variances seems to be an appropriate choice for the prior density. This prior appears to be more "balanced" than the uniform one. Since the support is the interval from zero to infinity, we should expect that our prior opinion is equally apportioned between points near zero and those near infinity. The prior $1/g$ does this as it bounds an infinite area over all regions of the form $(0,a)$ and (a,∞). The uniform prior puts infinite probability only on the latter region. That is, it seems biased towards larger values of the variance. In Section 3, some brief attention will be given to a proper prior, so those who have one can still employ the methods to be discussed.

All of the ideas presented in this section other than the use of normality are summarized in the following quote (Berger [2]):

> "We would indeed argue that noninformative prior Bayesian analysis is the *single most powerful method of statistical analysis,* in the sense of being the *ad hoc* method most likely to yield a sensible answer for a given investment of effort." (author's italics)

2. THE HIERARCHICAL NORMAL LINEAR MODEL

In this section, the algebraic manipulations required to evaluate the three level hierarchical model are performed. Attention will be restricted to linear versions of the model. This is done mostly for computational convenience.

Before beginning the manipulations, a few notational items will be presented. Scalars will be represented by lower case letters. Vectors will be represented by bold face characters. Matrices will be represented by upper case letters. In classical statistics it is common to use upper case symbols to represent random variables. In a Bayesian analysis the various quantities are sometimes random and are sometimes fixed, so no attempt is made to use notation to identify random quantities. For example, in the model, the data are random and the parameters are fixed, but in the posterior, the parameters are random and the data are fixed. At times, the distribution of some parameters conditioned on others is needed. When examining a density function, the way to tell the fixed quantities from the random ones is to look at the left hand side. For example, $p(\boldsymbol{\theta}|y,G,H)$ indicates that in the function which follows, $\boldsymbol{\theta}$ is the random quantity and is the variable in the density, while y, G, and H are fixed quantities. The density is for the indicated random variable, conditioned on the specific values given. When the two sides are separated by a proportionality symbol ($\propto$), the constant of proportionality may depend upon the conditional items. The constant can always be found by integrating the function with respect to the random elements.

2.1 The Model

The linear version of the three level Bayesian model is usually attributed to Lindley and Smith [24]. The three levels are:

$$y|\boldsymbol{\theta},G \sim N(A\boldsymbol{\theta},G) \qquad (N \times 1),$$

$$\boldsymbol{\theta}|\boldsymbol{\mu},H \sim N(B\boldsymbol{\mu},H) \qquad (k \times 1),$$

and

$$\boldsymbol{\mu} \sim N(\boldsymbol{\rho},C) \qquad (z \times 1),$$

where $\boldsymbol{\rho}$ and C are known and A and B (also known) are of full rank. A special case, and the only one considered here, is obtained by letting $C^{-1} \to 0$. This is equivalent to setting $p(\boldsymbol{\mu}) \propto \mathbf{1}$, the widely accepted noninformative prior for the mean. It is not necessary in this case to make any statement about $\boldsymbol{\rho}$. In most applications the covariance matrices G and H will not be known. It is then necessary to specify a prior distribution for them. Let $p(G,H)$ be the density for this prior distribution.

The standard credibility problem is to make inferences about $\boldsymbol{\theta}$, the expected losses (or loss ratios) for the various groups under consideration. The matrix A reflects the nature of the data collected. For example, there may be data from various years for each group. The second level indicates any relationships between the groups. One particular version of this model is analyzed in Sections 3 and 4; examples of other models are presented in Section 5. In any event, the objective of all the manipulations in this section is to obtain the posterior distribution and moments of $\boldsymbol{\theta}$ given the data y. Of less interest are the posterior distributions of G and H.

2.2 Three Helpful Mathematical Items

The first useful relationship is a matrix equation that is true for any symmetric non-singular matrix G; it will be used for completing the square.

$$x'Gx - 2x'B = (x - G^{-1}B)'G(x - G^{-1}B) - B'G^{-1}B.$$

The second item relates to the multivariate normal density. In general, the multivariate normal p.d.f. for a random variable with mean $\boldsymbol{\mu}$ and covariance C (a positive definite matrix) is

$$f(x) = (2\pi|C|)^{-1/2}\exp[-(x - \boldsymbol{\mu})\,'C^{-1}(x - \boldsymbol{\mu})/2],$$

where $|C|$ denotes the determinant of the matrix C. This implies that in general

$$\int\exp[-\,(x - \boldsymbol{\mu})\,'C^{-1}(x - \boldsymbol{\mu})/2]dx = (2\pi|C|)^{1/2}.$$

The final item is concerned with finding conditional densities. The general problem is the following: Let $f(a_1, \ldots a_m | b_1, \ldots, b_n)$ be proportional to the conditional density of $A_1, \ldots, A_m$ given $B_1 = b_1, \ldots, B_n = b_n$ (n may be 0). To find the conditional density of $A_1, \ldots, A_g$ given $B_1 = b_1, \ldots, B_h = b_h$ (where $g \leq m$ and $h \leq n$ and at least one of the inequalities is strict) evaluate $\int \ldots \int f(a_1, \ldots a_m | b_1, \ldots, b_n) da_{g+1} \ldots da_m$ and then drop all terms that involve only $b_1, \ldots, b_n$. (This latter step can be done first; additional terms can be eliminated after the integration.) The resulting function will be proportional to the desired density.

A related fact is that a conditional density is proportional to the conditional density in which some of the quantities on the left hand side of the "|" are moved to the right hand side. For example, $f(a_1, a_2 | b_1, b_2)$ is proportional to the conditional density of A_1 given $A_2 = a_2$, $B_1 = b_1$, $B_2 = b_2$. Any factors that depend only on a_2 can be deleted.

2.3 Two Useful, Non-Bayesian Quantities

In the first level of the model, with G assumed known and $\boldsymbol{\theta}$ taken to represent a fixed, but unknown, parameter, the classical least-squares estimator of $\boldsymbol{\theta}$ is found by minimizing

$$\begin{aligned}&(y - A\boldsymbol{\theta})'\, G^{-1}\, (y - A\boldsymbol{\theta})\\ &= y'\, G^{-1}y - 2\boldsymbol{\theta}'\, A'\, G^{-1}y + \boldsymbol{\theta}'\, A'\, G^{-1}A\boldsymbol{\theta}\\ &= y'\, G^{-1}y +\\ &\quad [\boldsymbol{\theta} - (A'\, G^{-1}A)^{-1}\, A'\, G^{-1}y]'\, (A'\, G^{-1}A)\, [\boldsymbol{\theta} - (A'\, G^{-1}A)^{-1}\, A'\, G^{-1}y].\end{aligned}$$

Let $\Lambda = (A'\, \mathrm{G}^{-1}A)^{-1}$, a positive definite matrix. Then the minimum must occur at $\hat{\boldsymbol{\theta}} = \Lambda A'\, G^{-1}y$.

Combining the first two levels gives

$$y | \boldsymbol{\mu}, G, H \sim N(AB\boldsymbol{\mu}, G + AHA').$$

The same manipulations yield

$$\hat{\boldsymbol{\mu}} = [B'A'(G + AHA')^{-1}AB]^{-1}B'A'(G + AHA')^{-1}y$$

and some matrix algebra produces the alternative form

$$\hat{\boldsymbol{\mu}} = [B'(H + \Lambda)^{-1}B]^{-1}B'(H + \Lambda)^{-1}\hat{\boldsymbol{\theta}}.$$

While the theory behind the above development is not germane to a Bayesian analysis, it is comforting to note that a Bayesian analysis often produces results that match those from classical theory. The quantities $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\mu}}$ will appear often in the analysis that follows, but will arise from a different algebraic procedure.

2.4 The Joint Density of $(y,\boldsymbol{\theta},\boldsymbol{\mu},G,H)$

The joint density involving all of the quantities from the three levels is the ideal place to begin. The last concept presented in Section 2.2 indicates that it is also the conditional density of any subset of the five variables given the remaining ones. The density is given by

$$p(y|\boldsymbol{\theta},G)p(\boldsymbol{\theta}|\boldsymbol{\mu},H)p(\boldsymbol{\mu})p(G,H)$$

which is proportional to (recalling that $p(\boldsymbol{\mu}) \propto \mathbf{1}$)

$$p(G,H)(2\pi)^{-(N+k)/2}(|G||H|)^{-1/2}$$
$$\times \exp[-(y - A\boldsymbol{\theta})'G^{-1}(y - A\boldsymbol{\theta})/2 - (\boldsymbol{\theta} - B\boldsymbol{\mu})'H^{-1}(\boldsymbol{\theta} - B\boldsymbol{\mu})/2].$$

It would be pleasant to proceed directly to the density of $\boldsymbol{\theta}$ given y. However, it is not possible to obtain this density analytically. Instead, begin by obtaining those conditional densities that are reasonably easy to derive. This is done in the next section. In the following section these densities are used to obtain the desired result.

2.5 Several Conditional Densities

The following subsections contain the derivations of a number of important conditional densities. The results are summarized in Section 2.5.6. Readers who are uninterested in the derivations can skip to that point.

2.5.1 Density of $\boldsymbol{\theta}|y,G,H$

This derivation is presented in great detail in order to indicate how these calculations are done. Since the joint density from Section 2.4 is also the conditional density of $(\boldsymbol{\theta},\boldsymbol{\mu}|y,G,H)$ it is only necessary to integrate $\boldsymbol{\mu}$ out of the joint density. Begin by removing terms involving only $y,G,$ $H,$ and constants. The remaining part of the joint density is

$$\exp[-(y - A\boldsymbol{\theta})'G^{-1}(y - A\boldsymbol{\theta})/2 - (\boldsymbol{\theta} - B\boldsymbol{\mu})'H^{-1}(\boldsymbol{\theta} - B\boldsymbol{\mu})/2]$$
$$= \exp[-(y - A\boldsymbol{\theta})'G^{-1}(y - A\boldsymbol{\theta})/2 - \boldsymbol{\theta}'H^{-1}\boldsymbol{\theta}/2 + 2\boldsymbol{\mu}'\mathrm{B}'\mathrm{H}^{-1}\boldsymbol{\theta}/2$$
$$- \boldsymbol{\mu}'B'H^{-1}B\boldsymbol{\mu}/2].$$

Let $\Xi = (B'H^{-1}B)^{-1}$. Completing the square with respect to $\boldsymbol{\mu}$ in the above expression yields

$$\exp[(-(y - A\boldsymbol{\theta})'G^{-1}(y - A\boldsymbol{\theta})/2 - \boldsymbol{\theta}'H^{-1}\boldsymbol{\theta}/2$$
$$- (\boldsymbol{\mu} - \Xi B'H^{-1}\boldsymbol{\theta})'\Xi^{-1}(\boldsymbol{\mu} - \Xi B'H^{-1}\boldsymbol{\theta})/2 + \boldsymbol{\theta}'H^{-1}B\Xi B'H^{-1}\boldsymbol{\theta}/2].$$

Now integrate with respect to $\boldsymbol{\mu}$. $\boldsymbol{\mu}$ only appears as the quadratic form in the third term and upon integration produces only constants and the determinant of Ξ. Since Ξ is a function of H only, it can be dropped. Therefore the density is proportional to

$$\exp[-(y - A\boldsymbol{\theta})'G^{-1}(y - A\boldsymbol{\theta})/2 - \boldsymbol{\theta}'H^{-1}\boldsymbol{\theta}/2 + \boldsymbol{\theta}'H^{-1}B\Xi B'H^{-1}\boldsymbol{\theta}/2].$$

Expand the first term and remove the part not involving $\boldsymbol{\theta}$. The result is now

$$\exp[-\boldsymbol{\theta}'(A'G^{-1}A - H^{-1}B\Xi B'H^{-1} + H^{-1})\boldsymbol{\theta}/2 + \boldsymbol{\theta}'A'G^{-1}y].$$

Let $V^{-1} = A'G^{-1}A - H^{-1}B\Xi B'H^{-1} + H^{-1}$.
Complete the square on $\boldsymbol{\theta}$ to obtain

$$\exp[-(\boldsymbol{\theta} - VA'G^{-1}y)'V^{-1}(\boldsymbol{\theta} - VA'G^{-1}y)/2 + y'G^{-1}AVA'G^{-1}y/2].$$

Since the second term does not depend on $\boldsymbol{\theta}$ it can be dropped. The final result is

$$p(\boldsymbol{\theta}|y,G,H) \propto \exp[-(\boldsymbol{\theta} - VA'G^{-1}y)'V^{-1}(\boldsymbol{\theta} - VA'G^{-1}y)/2].$$

By inspection it is immediately apparent that

$$\boldsymbol{\theta}|y,G,H \sim N(VA'G^{-1}y,V).$$

Let $\tilde{\boldsymbol{\theta}} = VA'G^{-1}y$ be the conditional mean. It can be rewritten as

$$\tilde{\boldsymbol{\theta}} = (H^{-1} + \Lambda^{-1})^{-1}(\Lambda^{-1}\hat{\boldsymbol{\theta}} + H^{-1}B\hat{\boldsymbol{\mu}}).$$

This is the customary weighted average common in a Bayesian analysis. It is also a (linear) credibility formula. In fact, this is the result that arises from an EB analysis. As discussed in Section 1, the point of departure is the treatment of the unknown G and H.

2.5.2 *Density of* $\boldsymbol{\mu}|y,G,H$

This calculation is included mostly for completeness. It is not used in any subsequent work. The procedure is exactly the same as that used above, only now integrate out $\boldsymbol{\theta}$ instead of $\boldsymbol{\mu}$. The result is

$$\boldsymbol{\mu}|y,G,H \sim N(\hat{\boldsymbol{\mu}},[B'(\Lambda + H)^{-1}B]^{-1}).$$

2.5.3 *Density of* $\boldsymbol{\theta},G,H|y$

To find this density, integrate $\boldsymbol{\mu}$ out of the joint density. The difference between this calculation and the one in Section 2.5.1 is that terms involving G

and H must be retained. The result is

$$\begin{aligned} p(\boldsymbol{\theta},G,H|y) \propto\ & p(G,H)|G|^{-1/2}|H|^{-1/2}|\Xi|^{1/2} \\ & \times \exp[-(y - A\boldsymbol{\theta})'G^{-1}(y - A\boldsymbol{\theta})/2 - \boldsymbol{\theta}'H^{-1}\boldsymbol{\theta}/2 \\ & + \boldsymbol{\theta}'H^{-1}B\Xi B'H^{-1}\boldsymbol{\theta}/2]. \end{aligned}$$

Let $Q = H^{-1} - H^{-1}B\Xi B'H^{-1}$. Then

$$\begin{aligned} p(\boldsymbol{\theta},G,H|y) \propto\ & p(G,H)|G|^{-1/2}|H|^{-1/2}|\Xi|^{1/2} \\ & \times \exp[-(y - A\boldsymbol{\theta})'G^{-1}(y - A\boldsymbol{\theta})/2 - \boldsymbol{\theta}'Q\boldsymbol{\theta}/2]. \end{aligned}$$

2.5.4 *Density of* $\boldsymbol{\mu},G,H|y$

Begin by integrating $\boldsymbol{\theta}$ out of the joint density, once again retaining terms involving G and H.

$$\begin{aligned} p(\boldsymbol{\mu},G,H|y) \propto\ & p(G,H)|G|^{-1/2}|H|^{-1/2}\int\exp[-yG^{-1}y/2 + 2\boldsymbol{\theta}A'G^{-1}y/2 \\ & - \boldsymbol{\theta}'A'G^{-1}A\boldsymbol{\theta}/2 - \boldsymbol{\theta}'H^{-1}\boldsymbol{\theta}/2 + 2\boldsymbol{\theta}'H^{-1}B\boldsymbol{\mu}/2 \\ & - \boldsymbol{\mu}'B'H^{-1}B\boldsymbol{\mu}/2]d\boldsymbol{\theta} \\ \propto\ & p(G,H)|G|^{-1/2}|H|^{-1/2}\int\exp[-yG^{-1}y/2 - \boldsymbol{\theta}'(\Lambda^{-1} + H^{-1})\boldsymbol{\theta}/2 \\ & + 2\boldsymbol{\theta}'(A'G^{-1}y + H^{-1}B\boldsymbol{\mu})/2 - \boldsymbol{\mu}'\Xi^{-1}\boldsymbol{\mu}/2]d\boldsymbol{\theta}. \end{aligned}$$

The third term in the exponent can be written $\Lambda^{-1}\hat{\boldsymbol{\theta}} + H^{-1}B\boldsymbol{\mu}$. Complete the square to obtain

$$\begin{aligned} p(\boldsymbol{\mu},G,H|y) \propto\ & p(G,H)|G|^{-1/2}|H|^{-1/2}\int\exp\{-y'G^{-1}y/2 \\ & - [\boldsymbol{\theta} - (\Lambda^{-1} + H^{-1})^{-1}(\Lambda^{-1}\hat{\boldsymbol{\theta}} + H^{-1}B\boldsymbol{\mu})]'(\Lambda^{-1} + H^{-1}) \\ & \times [\boldsymbol{\theta} - (\Lambda^{-1} + H^{-1})^{-1}(\Lambda^{-1}\hat{\boldsymbol{\theta}} + H^{-1}B\boldsymbol{\mu})] \\ & + (\Lambda^{-1}\hat{\boldsymbol{\theta}} + H^{-1}B\boldsymbol{\mu})'(\Lambda^{-1} + H^{-1})^{-1}(\Lambda^{-1}\hat{\boldsymbol{\theta}} + H^{-1}B\boldsymbol{\mu}) \\ & - \boldsymbol{\mu}'\Xi^{-1}\boldsymbol{\mu}/2\}d\boldsymbol{\theta} \\[1ex] \propto\ & p(G,H)|G|^{-1/2}|H|^{-1/2}|\Lambda^{-1} + H^{-1}|^{-1/2} \times \exp[-y'G^{-1}y/2 \\ & - \boldsymbol{\mu}'\Xi^{-1}\boldsymbol{\mu}/2 + \boldsymbol{\mu}'B'H^{-1}(\Lambda^{-1} + H^{-1})^{-1}H^{-1}B\boldsymbol{\mu}/2 \\ & + 2\boldsymbol{\mu}B'H^{-1}(\Lambda^{-1} + H^{-1})^{-1}\Lambda^{-1}\hat{\boldsymbol{\theta}}/2 \\ & + \hat{\boldsymbol{\theta}}'\Lambda^{-1}(\Lambda^{-1} + H^{-1})^{-1}\Lambda^{-1}\hat{\boldsymbol{\theta}}/2]. \end{aligned}$$

Now use the three identities

$$\begin{aligned} & H^{-1}(\Lambda^{-1} + H^{-1})^{-1}H^{-1} = H^{-1} - (\Lambda + H)^{-1}, \\ & H^{-1}(\Lambda^{-1} + H^{-1})^{-1}\Lambda^{-1} = (\Lambda + H)^{-1}, \text{ and} \\ & \Lambda^{-1}(\Lambda^{-1} + H^{-1})^{-1}\Lambda^{-1} = \Lambda^{-1} - (\Lambda + H)^{-1}. \end{aligned}$$

Let $\Psi = \Lambda + H$. Then,

$$\begin{aligned} p(\boldsymbol{\mu},G,H|y) \propto\ & p(G,H)|G|^{-1/2}|H|^{-1/2}|\Lambda^{-1} + H^{-1}|^{-1/2} \times \exp[-y'G^{-1}y/2 \\ & - \boldsymbol{\mu}'\Xi^{-1}\boldsymbol{\mu}/2 + \boldsymbol{\mu}'B'H^{-1}B\boldsymbol{\mu}/2 - \boldsymbol{\mu}'B'\Psi^{-1}B\boldsymbol{\mu}/2 \\ & + 2\boldsymbol{\mu}'B'\Psi^{-1}\hat{\boldsymbol{\theta}}/2 + \hat{\boldsymbol{\theta}}'\Lambda^{-1}\hat{\boldsymbol{\theta}}/2 - \hat{\boldsymbol{\theta}}'\Psi^{-1}\hat{\boldsymbol{\theta}}/2]. \end{aligned}$$

The second and third terms in the exponent cancel. From Section 2.3, $(B'\Psi^{-1}B)\hat{\boldsymbol{\mu}} = B'\Psi^{-1}\hat{\boldsymbol{\theta}}$. Complete the square to obtain

$$\begin{aligned} p(\boldsymbol{\mu},G,H|y) \propto\ & p(G,H)|G|^{-1/2}|H|^{-1/2}|\Lambda^{-1} + H^{-1}|^{-1/2} \times \exp[-y'G^{-1}y/2 \\ & - (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})'(B'\Psi^{-1}B)(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})/2 + \hat{\boldsymbol{\mu}}'(B'\Psi^{-1}B)\hat{\boldsymbol{\mu}}/2 \\ & + \hat{\boldsymbol{\theta}}'\Lambda^{-1}\hat{\boldsymbol{\theta}}/2 - \hat{\boldsymbol{\theta}}'\Psi^{-1}\hat{\boldsymbol{\theta}}/2]. \end{aligned}$$

2.5.5 *Density of G,H|y*

Integrate $\boldsymbol{\mu}$ out of the density in Section 2.5.4 to obtain

$$\begin{aligned} p(G,H|y) \propto\ & p(G,H)|G|^{-1/2}|H|^{-1/2}|\Lambda^{-1} + H^{-1}|^{-1/2}|B'\Psi^{-1}B|^{-1/2} \\ & \times \exp[-y'G^{-1}y/2 + \hat{\boldsymbol{\theta}}'\Lambda^{-1}\hat{\boldsymbol{\theta}}/2 - \hat{\boldsymbol{\theta}}'\Psi^{-1}\hat{\boldsymbol{\theta}}/2 + \hat{\boldsymbol{\mu}}'(B'\Psi^{-1}B)\hat{\boldsymbol{\mu}}/2]. \end{aligned}$$

In the second term of the exponent write $\hat{\boldsymbol{\theta}}$ in terms of y. Then complete the squares to obtain

$$\begin{aligned} p(G,H|y) \propto\ & p(G,H)|G|^{-1/2}|H|^{-1/2}|\Lambda^{-1} + H^{-1}|^{-1/2}|B'\Psi^{-1}B|^{-1/2} \\ & \times \exp[-(y - A\hat{\boldsymbol{\theta}})'G^{-1}(y - A\hat{\boldsymbol{\theta}})/2 \\ & - (\hat{\boldsymbol{\theta}} - B\hat{\boldsymbol{\mu}})'\Psi^{-1}(\hat{\boldsymbol{\theta}} - B\hat{\boldsymbol{\mu}})/2]. \end{aligned}$$

The two terms in the exponent are the within and between sums of squares, respectively. Both depend on the unknown variances, G and H. While it is once again comforting to note that frequentist quantities have appeared in the Bayesian development, we should keep in mind that these quantities have no special meaning. The determinants can be rewritten as

$$(|G^{-1}|/|A'G^{-1}A|)^{1/2}(|\Psi^{-1}|/|B'\Psi^{-1}B|)^{1/2}.$$

The two numerator terms can form the basis for a prior distribution on (G,H). This is somewhat consistent with the ideas presented in Box and Tiao [3] and in Tiao and Zellner [33].

2.5.6 *Summary*

The important matrices and distributions from this section are repeated for convenience:

$$\Lambda = (A'G^{-1}A)^{-1}$$

$$\Xi = (B'H^{-1}B)^{-1}$$

$$\Psi = \Lambda + H$$

$$Q = H^{-1} - H^{-1}B\Xi B'H^{-1}$$

$$V = [A'G^{-1}A - H^{-1}B\Xi B'H^{-1} + H^{-1}]^{-1} = (\Lambda^{-1} + Q)^{-1}$$

$$\hat{\boldsymbol{\theta}} = \Lambda A'G^{-1}y$$

$$\hat{\boldsymbol{\mu}} = [B'(H + \Lambda)^{-1}B]^{-1}B'(H + \Lambda)^{-1}\hat{\boldsymbol{\theta}}$$

$$\tilde{\boldsymbol{\theta}} = (H^{-1} + \Lambda^{-1})^{-1}(\Lambda^{-1}\hat{\boldsymbol{\theta}} + H^{-1}B\hat{\boldsymbol{\mu}})$$

$$\boldsymbol{\theta}|y,G,H \sim N(\tilde{\boldsymbol{\theta}},V)$$

$$\begin{aligned} p(\boldsymbol{\theta},G,H|y) \propto\ & p(G,H)|G|^{-1/2}|H|^{-1/2}|\Xi|^{1/2} \\ & \times \exp[-(y - A\boldsymbol{\theta})'G^{-1}(y - A\boldsymbol{\theta})/2 - \boldsymbol{\theta}'Q\boldsymbol{\theta}/2] \end{aligned}$$

$$\begin{aligned} p(G,H|y) \propto\ & p(G,H)|G|^{-1/2}|\Lambda|^{1/2}|\Psi|^{-1/2}|B'\Psi^{-1}B|^{-1/2} \\ & \times \exp[-(y - A\hat{\boldsymbol{\theta}})'G^{-1}(y - A\hat{\boldsymbol{\theta}})/2 \\ & - (\hat{\boldsymbol{\theta}} - B\hat{\boldsymbol{\mu}})'\Psi^{-1}(\hat{\boldsymbol{\theta}} - B\hat{\boldsymbol{\mu}})/2]. \end{aligned}$$

2.6 Two Empirical Bayes Approaches to Estimating $\boldsymbol{\theta}$

As introduced in Section 1.2.2, the EB approach begins by finding the posterior mean of $\boldsymbol{\theta}$ given the covariance terms G and H. In the HNLM this is $\tilde{\boldsymbol{\theta}}$. External estimates are then found for G and H. In this section, two general approaches to finding such estimates are introduced.

The first method uses the posterior density $p(G,H|y)$. Either the mean or the mode could be used as the estimate. The mean is superior in that it is guaranteed to be in the interior of the parameter space. The mode is often easier to compute, but may be on a boundary. Either estimate usually requires a numerical evaluation. Specific formulas for a simple model are presented in Section 3.8.

The second method is an iterative technique. Begin with a preliminary estimate of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$. Then in $p(\boldsymbol{\theta},G,H|y)$ hold $\boldsymbol{\theta}$ fixed at its current value and find the values of G and H that maximize this density. Obtain a revised estimate of $\boldsymbol{\theta}$ by evaluating $\tilde{\boldsymbol{\theta}}$ at the values of G and H just obtained. Repeat this procedure until $\tilde{\boldsymbol{\theta}},G$, and H stabilize. It is not entirely clear what the results mean, but the procedure is similar to that recommended by Morris [27]. Computationally, this tends to be the simplest approach, as the maximization can often be done analytically. This is demonstrated for a simple model in Section 3.5. If an analytical approach is not possible, an all-purpose maximization method like that of Nelder and Mead [29] is likely to provide the answer.

Recall that one of the drawbacks of the usual EB method is its inability to produce variance estimates. In Section 3.8, it is shown how this can be done when G and H are estimated by the posterior mean.

2.7 Finding Posterior Quantities by Integration

All of the integrations done up to this point were easy to accomplish by completing the square. The one needed to obtain the posterior density of $\boldsymbol{\theta}|y$ is found from

$$p(\boldsymbol{\theta}|y) = \int p(\boldsymbol{\theta}|G,H,y)p(G,H|y)dGdH.$$

The first density in the integrand is a multivariate normal density and was obtained in Section 2.5.1. The second density was obtained in Section 2.5.5 up to a constant of proportionality. We must obtain that constant in order to insert the exact density in the integral above. That can be found by integrating the expression found in Section 2.5.5 with respect to both G and H. These two integrals are of equal difficulty and usually must be done numerically. The degree of difficulty will depend on the form of the covariance matrices G and H and the prior density $p(G,H)$. It will be seen in Section 3 that in a specific case the problem can be analytically reduced to a one-dimensional numerical integration. Some excellent procedures for performing multidimensional numerical integration are given in Smith, Skene, Shaw, Naylor, and Dransfield [32].

In most applications, the vector $\boldsymbol{\theta}$ will be of a reasonably high dimension, certainly greater than two. It is unlikely that much insight will be gained by examining the posterior density. The remainder of this section is devoted to obtaining various summary quantities. This will conclude the development of the general hierarchical normal linear model.

2.7.1 Posterior Mean of $\theta_i|y$

One way to obtain this quantity would be to evaluate the following integral:

$$\int \theta_i p(\boldsymbol{\theta}|y)d\boldsymbol{\theta}.$$

Given the fact that a numerical step is necessary to yield each evaluation of the integrand, the cost of performing this integration is likely to be quite high. Instead, employ the following result (this notion was introduced in Section 1.2.3):

$$\mathrm{E}(\theta_i|y) = \mathrm{E}[\mathrm{E}(\theta_i|G,H,y)] = \mathrm{E}(\tilde{\theta}_i|y) = \int \tilde{\theta}_i p(G,H|y)dGdH.$$

Note that $\tilde{\theta}_i$ is a function of G and H.

2.7.2 *Posterior Variance of* $\theta_i|y$

A similar argument yields the following.

$$\begin{aligned}\mathrm{Var}(\theta_i|y) &= \mathrm{Var}[\mathrm{E}(\theta_i|G,H,y)] + \mathrm{E}[\mathrm{Var}(\theta_i|G,H,y)]\\ &= \mathrm{Var}(\tilde{\theta}_i|y) + \mathrm{E}(v_{ii}|y)\\ &= \int(\tilde{\theta}_i)^2 p(G,H|y)dGdH - [\mathrm{E}(\theta_i|y)]^2 + \int v_{ii}p(G,H|y)dGdH\end{aligned}$$

where v_{ii} is the i^{th} diagonal element of the matrix V introduced in Section 2.5.1.

2.7.3 *Posterior Density of* $\theta_i|y$

This univariate density could be plotted to provide insight about a particular group mean. An approximate integration needs to be performed to get each point from the posterior density. The formula is

$$p(\theta_i|y) = \int p(\theta_i|G,H,y)p(G,H|y)dGdH.$$

The first density is a univariate normal density with mean $\tilde{\theta}_i$ and variance v_{ii}.

If this calculation appears to be too time-consuming, the posterior distribution may be approximated by a normal distribution with moments as given in Sections 2.7.1 and 2.7.2. This result is given in Berger [2] and is a Bayesian version of the central limit theorem. The same result applies in the following section.

2.7.4 *Predictive Density of a Future Observation*

In the insurance setting it may be more useful to get information about the losses in a future period than to estimate the class mean. Such a calculation would incorporate both the uncertainty with respect to the group mean and the uncertainty about the experience of next year's insureds.

In general, consider a new observation, $\boldsymbol{x} \sim N(A_x\boldsymbol{\theta},C_x)$ where C_x will depend in some way on the elements of G. A typical example would have A_x be a $1 \times k$ vector of zeros with a one in the i^{th} column. This would make x (a scalar) an observation from the i^{th} group. The matrix C_x would be a scalar of the form σ^2/P where σ^2 is the variance from the original model and P is a measure of exposure for the year to come.

The density of interest is

$$p(\boldsymbol{x}|y) = \int p(\boldsymbol{x}|\boldsymbol{\theta},G,H,y)p(\boldsymbol{\theta},G,H|y)d\boldsymbol{\theta}dGdH.$$

This is likely to be difficult to obtain. It is much easier to get the moments.

$$\begin{aligned} \mathrm{E}[\boldsymbol{x}|\boldsymbol{y}] &= \mathrm{E}[\mathrm{E}(\boldsymbol{x}|\boldsymbol{\theta},G,H,\boldsymbol{y})] \\ &= \mathrm{E}[A_x\boldsymbol{\theta}|\boldsymbol{y}] = A_x\mathrm{E}[\boldsymbol{\theta}|\boldsymbol{y}]. \end{aligned}$$

The expectation can be found using the formula in Section 2.7.1 since each element of the vector of expectations can be found individually.

For the variance

$$\begin{aligned} \mathrm{Cov}[\boldsymbol{x}|\boldsymbol{y}] &= \mathrm{Cov}[\mathrm{E}(\boldsymbol{x}|\boldsymbol{\theta},G,H,\boldsymbol{y})] + \mathrm{E}[\mathrm{Cov}(\boldsymbol{x}|\boldsymbol{\theta},G,H,\boldsymbol{y})] \\ &= \mathrm{Cov}[A_x\boldsymbol{\theta}|\boldsymbol{y}] + \mathrm{E}[C_x|\boldsymbol{y}] \\ &= A_x\mathrm{Cov}[\boldsymbol{\theta}|\boldsymbol{y}]A_x' + \mathrm{E}[C_x|\boldsymbol{y}]. \end{aligned}$$

The covariance requires evaluation of $\mathrm{Cov}(\theta_i,\theta_j|\boldsymbol{y})$. This can be done from $\int v_{ij}p(G,H|\boldsymbol{y})dGdH$. The $(ij)^{\text{th}}$ term of the expected value is evaluated as $\int (C_x)_{ij}p(G,H|\boldsymbol{y})dGdH$.

3. THE ONE-WAY MODEL

In this section a specific hierarchical model is investigated. It is similar to the model treated by Bühlmann and Straub [7] in their EB analysis and is appropriate when there are k identically distributed groups and the goal is the simultaneous estimation of their means. The three levels of the one-way model are

Level 1—$y_{ij}|\theta_i,\sigma^2 \sim N(\theta_i,\sigma^2/P_{ij})$ $i = 1, \ldots, k \quad j = 1, \ldots, n_i$

Level 2—$\theta_i|\mu,\tau^2 \sim N(\mu,\tau^2)$

Level 3—$\mu \sim N(0,\infty)$.

The random variables at each level are conditionally independent and P_{ij} is some measure of exposure. The usual situation is that y_{ij} is the average loss (or loss ratio) in year j for class i. This differs from the Bühlmann-Straub model in just one respect. In their model, the level one variances were allowed to differ from class to class. Their result, however, uses only the average of these variances. That is, at no point is this variability taken into account. An indication of how one could truly account for unequal variances is given in Section 5.

To use the formulas of the previous section it is necessary to identify the various matrices and vectors. Begin by letting $\boldsymbol{y}$ be the $N \times 1$ vector of the observations where $N = \Sigma n_i$. Arrange the observations so $y_{11}, \ldots, y_{1n_1}$ appears first, followed by $y_{21}, \ldots, y_{2n_2}$, and so forth. The matrix A is $N \times k$ and contains only zeros and ones. In the first column, the ones are in the first n_1 rows. In the second column, the ones are in rows $n_1 + 1$ through $n_1 + n_2$, and so forth. The vector $\boldsymbol{\theta}$ is $k \times 1$ and contains the unknown group means, $\theta_1, \ldots, \theta_k$. The covariance matrix G is diagonal with diagonal elements running from σ^2/P_{11} in the upper left corner to σ^2/P_{kn_k} in the lower right corner. At the second level, B is a $k \times 1$ vector consisting entirely of ones. Let $\mathbf{1}$ indicate such a vector. The vector $\boldsymbol{\mu}$ is a scalar and so will be written μ. The covariance matrix H is diagonal with all elements equal to τ^2, that is, $H = \tau^2 I_k$.

The exposition will proceed in four steps. The first is a development of a pair of useful matrix relationships. They will aid in the evaluation of the determinants and inverses. Next, prior distributions for σ^2 and τ^2 are introduced. The third step is to obtain the conditional densities. The final step is to perform the integrations.

3.1 Two Useful Matrix Facts

If A is a nonsingular matrix and $\boldsymbol{c}$ and $\boldsymbol{d}$ are vectors, then

$$|A + \boldsymbol{cd}'| = |A|(1 + \boldsymbol{d}'A^{-1}\boldsymbol{c}).$$

If the determinant is non-zero then

$$(A + \boldsymbol{cd}')^{-1} = A^{-1} - (A^{-1}\boldsymbol{cd}'A^{-1})/(1 + \boldsymbol{d}'A^{-1}\boldsymbol{c}).$$

For the special case where A is diagonal $(a_1, \ldots, a_k)$, $\boldsymbol{c} = c\mathbf{1}$, and $\boldsymbol{d} = \mathbf{1}$, the results are (where $J_k = \mathbf{11}'$, a $k \times k$ matrix)

$$|A + cJ_k| = (\Pi a_i)(1 + c\Sigma a_i^{-1})$$

and

$$(A + cJ_k)^{-1} \text{ has } (ii)^{\text{th}} \text{ term } a_i^{-1} - c/(a_i^2 b) \text{ and } (ij)^{\text{th}} \text{ term } -c/(a_i a_j b)$$

where $b = 1 + c\Sigma a_i^{-1}$.

Derivations of these results can be found in Graybill [12] (Theorem 8.9.3).

3.2 Prior Densities for (G,H)

In this section, three noninformative priors and one proper prior will be introduced. Two of the noninformative priors are based on a general theory that can be used in any setting of the HNLM. The third one is particular to the one-way model.

The easiest one to describe is the naive version of a noninformative prior. It is $p(G,H) \propto 1$. In the one-way model the only random elements are σ^2 and τ^2, so the actual prior in this case is $p(\sigma^2,\tau^2) \propto 1$. This prior is used by Morris [28]. While it is convenient for computational purposes, there are good theoretical reasons (Box and Tiao [3]) for not using it. The essence of the argument is that this prior puts too much weight on large values of the parameters. On the other hand, it is often the case that this improper prior will yield a proper posterior (something that must always be checked when using a noninformative prior). Also, the posterior mode is the maximum likelihood estimate. This should give comfort to those who are troubled by Bayesian methods. Call this prior 1.

The second prior is based on the arguments of Box and Tiao [3] for the balanced model. The balanced model is the special case where $n_1 = \ldots = n_k$ and $P_{11} = \ldots = P_{kn_k}$. The first requirement is common in insurance studies, as n_i often is the number of years of observation for the i^{th} class. However, it is extremely unlikely that the exposures will be equal for all years and all classes. In any event, in the one-way balanced model the reasonable noninformative prior is $p(\sigma^2,\tau^2) \propto (\sigma^2)^{-1}(\sigma^2 + Pn\tau^2)^{-1}$ where P is the common value of the P_{ij} and n is the common value of the n_i. A generalization for the unbalanced model is to use $p(\sigma^2,\tau^2) \propto (\sigma^2)^{-1}(\sigma^2 + m\tau^2)^{-1}$. Since m is to play the role of Pn one choice is $m = \Sigma P_{ij}/k$. An alternative is taken from the constant used when creating the unbiased frequentist estimator of τ^2. It is $m = [(\Sigma P_{ij})^2 - \Sigma(P_i)^2]/(k - 1)\Sigma P_{ij}$ where $P_i = \Sigma P_{ij}$. Call this (with arbitrary m) prior 2.

The final noninformative prior is, like the first one, available in all situations. In general, it is

$$p(G,H) \propto |G|^{-p\dim(G)/\dim(G)}|\Lambda + H|^{-p\dim(H)/\dim(H)}.$$

This prior is taken after Box and Tiao [3] and is related to the Fisher information about G and H. In the above expression, $p\dim(G)$ refers to the number of distinct parameters in the matrix G while $\dim(G)$ is the number of rows in G. For the one-way model, $p(\sigma^2,\tau^2) \propto (\sigma^2)^{-1}[\Pi(\sigma^2 + P_i\tau^2)]^{-1/k}$. In the balanced case this prior is identical to prior 2. Call it prior 3.

The fourth and final prior is an attempt to offer a proper distribution. As such, it requires that the investigator have a genuine opinion about the variances. When seeking a proper prior, mathematical convenience is always a high priority. At the very least, the family of proper priors should include a sufficiently large variety of possibilities so as to give the investigator a chance of finding a representative prior. The natural choice for variances is the inverse gamma distribution. The general form of the density is

$$p(x) \propto x^{-\nu}\exp(-\lambda/x).$$

For it to be a proper distribution we must have $\nu > 1$ and $\lambda > 0$. The limiting case of $\nu = 1$ and $\lambda = 0$ is similar to prior 2. Since prior 1 is equivalent to $\nu = 0$ and $\lambda = 0$ we see how far from being a proper distribution prior 1 is. The inverse gamma prior will be referred to as prior 4. A specific version appropriate for the one-way model will be given later.

3.3 Conditional Densities in the One-Way Model

This section contains all the details of the evaluation of the formulas in Section 2 in the special case of the one-way model.

3.3.1 Preliminary Quantities

$\Lambda = \text{diagonal } (\sigma^2/P_1, \ldots, \sigma^2/P_k)$ where $P_i = \Sigma_j P_{ij}$

$\Xi = \tau^2/k$ (a scalar)

$\hat{\theta}_i = \Sigma_j P_{ij} y_{ij}/P_i$

$\hat{\mu} = \Sigma w_i \hat{\theta}_i / w.$ where $w_i = P_i\tau^2/(\sigma^2 + P_i\tau^2)$ and $w. = \Sigma_i w_i$.

3.3.2 $\boldsymbol{\theta}|\boldsymbol{y},G,H$

This is a multivariate normal random variable. The mean vector $\tilde{\boldsymbol{\theta}}$ has i^{th} element

$$\tilde{\theta}_i = w_i\hat{\theta}_i + (1 - w_i)\hat{\mu}.$$

The matrix V^{-1} is

$$\begin{aligned} V^{-1} &= \text{diagonal } (\tau^{-2} + \sigma^{-2}P_i) - (k\tau^2)^{-1}J_k \\ &= \text{diagonal } [\tau^{-2}(1 - w_i)^{-1}] - (k\tau^2)^{-1}J_k, \end{aligned}$$

where as before J_k is a $k \times k$ matrix of 1's. The covariance matrix is V. Using the inversion formula from Section 3.1,

$$v_{ii} = \text{Var}(\theta_i|\boldsymbol{y},\sigma^2,\tau^2) = \tau^2(1 - w_i)[1 + (1 - w_i)/w.] \text{ and}$$

$$v_{ij} = \text{Cov}(\theta_i,\theta_j|\boldsymbol{y},\sigma^2,\tau^2) = \tau^2(1 - w_i)(1 - w_j)/w. \,.$$

3.3.3 $\boldsymbol{\theta},G,H|\boldsymbol{y}$

The two required sums of squares are

$$\begin{aligned}(y - A\boldsymbol{\theta})'G^{-1}(y - A\boldsymbol{\theta}) &= \Sigma_{ij}P_{ij}(y_{ij} - \theta_i)^2/\sigma^2\\ &= [\Sigma_{ij}P_{ij}(y_{ij})^2 - \Sigma_i P_i(\hat{\theta}_i)^2 + \Sigma_i P_i(\theta_i - \hat{\theta}_i)^2]/\sigma^2\end{aligned}$$

and

$$\boldsymbol{\theta}'Q\boldsymbol{\theta} = [\Sigma_i(\theta_i)^2 - k\bar{\theta}^2]/\tau^2 \text{ where } \bar{\theta} = \Sigma_i\theta_i/k.$$

The second version of the first quantity is useful for computational purposes as the first two sums depend only on the data while the last sum has only k terms. The desired density is

$$\begin{aligned}p(\boldsymbol{\theta},\sigma^2,\tau^2|y) \propto\ & p(\sigma^2,\tau^2)(\sigma^2)^{-N/2}(\tau^2)^{-(k-1)/2}\\ &\times \exp\{-\Sigma_{ij}P_{ij}(y_{ij} - \theta_i)^2/2\sigma^2 - (\Sigma_i(\theta_i)^2 - k\bar{\theta}^2)/2\tau^2\}.\end{aligned}$$

3.3.4 $G,H|\boldsymbol{y}$

Two important matrices are

$$\Psi = \text{diag}(\sigma^2/P_i + \tau^2) = \tau^2\text{diag}(1/w_i)$$

and

$$B'\Psi^{-1}B = w./\tau^2.$$

The two sums of squares are

$$(y - A\hat{\boldsymbol{\theta}})'G^{-1}(y - A\hat{\boldsymbol{\theta}}) = \Sigma_{ij}P_{ij}(y_{ij} - \hat{\theta}_i)^2/\sigma^2$$

and

$$(\hat{\boldsymbol{\theta}} - B\hat{\mu})'\Psi^{-1}(\hat{\boldsymbol{\theta}} - B\hat{\mu}) = \Sigma_i w_i(\hat{\theta}_i - \hat{\mu})^2/\tau^2.$$

The desired density is

$$\begin{aligned}p(\sigma^2,\tau^2|y) \propto\ & p(\sigma^2,\tau^2)(\sigma^2)^{-(N-k)/2}(\tau^2)^{-(k-1)/2}[\Pi_i w_i/w.]^{1/2}\\ &\times \exp[-\Sigma_{ij}P_{ij}(y_{ij} - \hat{\theta}_i)^2/2\sigma^2 - \Sigma_i(\hat{\theta}_i - \hat{\mu})^2 w_i/2\tau^2].\end{aligned}$$

For computation the exponent can be written

$$-[\Sigma_{ij}P_{ij}(y_{ij})^2 - \Sigma_i P_i(\hat{\theta}_i)^2]/2\sigma^2 - [\Sigma_i w_i(\hat{\theta}_i)^2 - (\Sigma_i w_i\hat{\theta}_i)^2/w.]/2\tau^2.$$

3.3.5 Prior Distributions for (σ^2,τ^2)

To make this section complete, the four priors developed in Section 3.2 are displayed.

Prior 1—$p(\sigma^2,\tau^2) \propto 1$.

Prior 2—$p(\sigma^2,\tau^2) \propto (\sigma^2)^{-1}(\sigma^2 + m\tau^2)^{-1}$.

Prior 3—$p(\sigma^2,\tau^2) \propto (\sigma^2)^{-1}[\Pi_i(\sigma^2 + P_i\tau^2)]^{-1/k}$.

Prior 4—$p(\sigma^2,\tau^2) \propto (\sigma^2)^{-\nu_1}(\tau^2)^{-\nu_2}\exp(-\lambda_1/\sigma^2 - \lambda_2/\tau^2)$.

Prior 4 uses two independent inverse chi-square random variables. Prior 1 is a special case. In the next section the four priors will be written with one general formula.

3.4 A Transformation

It turns out that calculations are much easier to carry out with a transformation of σ^2 and τ^2. The one to use is

$$\delta = \tau^2/\sigma^2 \text{ and } \alpha = \sigma^2.$$

The Jacobian for this transformation is α. The four prior densities become

Prior 1—$p(\alpha,\delta) \propto \alpha$.

Prior 2—$p(\alpha,\delta) \propto \alpha^{-1}(1 + m\delta)^{-1}$.

Prior 3—$p(\alpha,\delta) \propto \alpha^{-1}[\Pi_i(1 + P_i\delta)]^{-1/k}$.

Prior 4—$p(\alpha,\delta) \propto \alpha^{-(\nu_1+\nu_2-1)}\delta^{-\nu_2}\exp(-\lambda_1/\alpha - \lambda_2/\alpha\delta)$.

A general form that includes all four is

$$p(\alpha,\delta) \propto \alpha^{-q/2}h(\delta)\exp(-\lambda_1/\alpha - \lambda_2/\alpha\delta).$$

The two important conditional densities become (note that since the Jacobian was included in the prior, no other adjustments are needed)

$$\begin{aligned} p(\boldsymbol{\theta},\alpha,\delta|y) \propto\ & (\alpha)^{-(N+k+q-1)/2}(\delta)^{-(k-1)/2}h(\delta) \\ & \times \exp\{-[\lambda_1 + \Sigma_{ij}P_{ij}(y_{ij} - \theta_i)^2]/2\alpha \\ & - [\lambda_2 + (\Sigma_i(\theta_i)^2 - k\bar{\theta}^2)]/2\alpha\delta\} \end{aligned}$$

and

$$\begin{aligned} p(\alpha,\delta|y) \propto\ & (\alpha)^{-(N+q-1)/2}(\delta)^{-(k-1)/2}h(\delta)\ [\Pi_i w_i/w.]^{1/2} \\ & \times \exp\{-[\lambda_1 + \Sigma_{ij}P_{ij}(y_{ij} - \hat{\theta}_i)^2]/2\alpha \\ & - [\lambda_2 + \Sigma_i(\hat{\theta}_i - \hat{\mu})^2 w_i]/2\alpha\delta\}. \end{aligned}$$

The other important quantity is the distribution of $\boldsymbol{\theta}|\boldsymbol{y},\alpha,\delta$. From Section 3.3.2, it is multivariate normal. The i^{th} element of the mean vector is

$$\tilde{\theta}_i = w_i\hat{\theta}_i + (1 - w_i)\hat{\mu} \text{ where } w_i = P_i\delta/(1 + P_i\delta) \text{ and } \hat{\mu} = \Sigma_i w_i\hat{\theta}_i/w. \, .$$

The covariance matrix has diagonal elements

$$v_{ii} = \alpha\delta(1 - w_i)[1 + (1 - w_i)/w.]$$

and off diagonal elements

$$v_{ij} = \alpha\delta(1 - w_i)(1 - w_j)/w..$$

3.5 Iterative EB Estimates

Recall the two-step iterative procedure from Section 2.6. For the one-way model, the first step is to find the values of α and δ that maximize $p(\boldsymbol{\theta},\alpha,\delta|\boldsymbol{y})$ with $\tilde{\boldsymbol{\theta}}$ replacing $\boldsymbol{\theta}$. This density was obtained in the previous section. Let

$$C = \lambda_1 + \Sigma_{ij}P_{ij}(y_{ij} - \tilde{\theta}_i)^2$$

and

$$D = \lambda_2 + \Sigma_i(\tilde{\theta}_i - \tilde{\mu})^2 \text{ where } \tilde{\mu} = \Sigma_i\tilde{\theta}_i/k.$$

Differentiating the density with respect to α and δ produces the two equations

$$C + D/\delta - (N + k + q - 1)\alpha = 0$$

and

$$Dh(\delta)/\alpha - (k - 1)h(\delta)\delta + 2h'(\delta)\delta^2 = 0.$$

Solve the first equation for $\alpha = (C\delta + D)/(N + k + q - 1)\delta$. Insert this in the second equation to obtain

$$2Ch'(\delta)\delta^2 + [2Dh'(\delta) - (k - 1)Ch(\delta)]\delta + (N + q)Dh(\delta).$$

The solutions for the four priors are

Prior 1—$\delta = (N - 2)D/(k - 1)C$.

Prior 2—δ is the non-negative root of $(k + 1)mC\delta^2 + [(k - 1)C - NmD]\delta - (N + 2)D$. There is exactly one non-negative root and it is slightly larger than $ND/(k + 1)C$.

Prior 3—This one must be solved numerically.

Prior 4—$\delta = (N + 2\nu_1 - 2)D/(k - 1)C$.

The second step is to obtain the revised estimate of $\boldsymbol{\theta}$. It is

$$\tilde{\boldsymbol{\theta}} = w_i\hat{\boldsymbol{\theta}} + (1 - w_i)\hat{\boldsymbol{\mu}} \text{ where } w_i = P_i\delta/(1 + P_i\delta).$$

A surprising observation is that for prior 4 the estimate does not depend on ν_2. The solution with prior 1 is the ratio of the appropriate sums of squares but, unlike the usual EB estimate, it can never be negative. It is, however, possible to get a value of zero.

3.6 Density of $\delta|\boldsymbol{y}$

As indicated in Section 2.6, the iterative algorithm does not provide the other quantities of interest. Before developing the integration formulas, analytically integrate α out of the posterior density of $\alpha,\delta|\boldsymbol{y}$. For future use, let c be the constant of proportionality in the posterior density of $\alpha,\delta|\boldsymbol{y}$.

$$\begin{aligned} p(\delta|\boldsymbol{y}) = {} & \int c(\alpha)^{-(N+q-1)/2}(\delta)^{-(k-1)/2}h(\delta)[\Pi_i w_i/w.]^{1/2} \\ & \times \exp\{-[\lambda_1 + \Sigma_{ij}P_{ij}(y_{ij} - \hat{\theta}_i)^2]/2\alpha \\ & - [\lambda_2 + \Sigma_i(\hat{\theta}_i - \hat{\mu})^2 w_i]/2\alpha\delta\}d\alpha \\ = {} & c(\delta)^{-(k-1)/2}h(\delta)[\Pi_i w_i/w.]^{1/2} \times \{\lambda_1 + \Sigma_{ij}P_{ij}(y_{ij} - \hat{\theta}_i)^2 \\ & + [\lambda_2 + \Sigma_i(\hat{\theta}_i - \hat{\mu})^2 w_i]/\delta\}^{-(N+q-3)/2} \\ & \times \Gamma[(N + q - 3)/2]2^{(N+q-3)/2}. \end{aligned}$$

3.7 Evaluation by Integration

Let $f(\delta)$ be the essential part of $p(\delta|\boldsymbol{y})$. That is,

$$\begin{aligned} f(\delta) = {} & (\delta)^{-(k-1)/2}h(\delta)[\Pi_i w_i/w.]^{1/2} \\ & \times \{\lambda_1 + \Sigma_{ij}P_{ij}(y_{ij} - \hat{\theta}_i)^2 + [\lambda_2 + \Sigma_i(\hat{\theta}_i - \hat{\mu})^2 w_i]/\delta\}^{-(N+q-3)/2}. \end{aligned}$$

Let the constant $g = \int f(\delta)d\delta$. So $p(\delta|\boldsymbol{y}) = gf(\delta)$ and the relationship between the constants g and c is

$$c = \{g\Gamma[(N + q - 3)/2]2^{(N+q-3)/2}\}^{-1}.$$

The integral for g must be done numerically. An approach that is not necessarily the most efficient but is sure to work is to use separate numerical integrations on the intervals [0,1], [1,2], [2,4], [4,8], . . . until the contribution from the latest interval is sufficiently small. An iterative Gaussian integration converges fairly quickly. Any numerical analysis text (e.g., Burden, Faires, and Reynolds [8]) is likely to prove useful.

When doing numerical integration, it is important to know in advance if the integral will be finite. For the integral above it is sufficient to look at the balanced case. After removing some constants that depend only upon n and P, the integrand becomes

$$h(\delta)(1 + nP\delta)^{-(k-1)/2}[c_1 + \lambda_2/\delta + c_2/(1 + nP\delta)]^{-(N+q-3)/2}$$

with c_1 and c_2 being positive constants that depend only on the data. The four priors can be generalized to $h(\delta) = \delta^{-\nu_2}(1 + nP\delta)^{-h}$ where $\nu_2 = \nu_2$ for prior 4 and is zero otherwise and h is 1 for priors 2 and 3 and zero otherwise. It is necessary to verify the conditions under which the integral will exist as both $\delta \rightarrow 0$ and $\delta \rightarrow \infty$. For the first case, the essential part of the integrand is

$$\delta^{-\nu_2}[c_3 + \lambda_2/\delta]^{-(N+q-3)/2}.$$

For existence, either

$$\lambda_2 > 0 \text{ and } N + q - 2\nu_2 > 1 \text{ or}$$

$$\lambda_2 = 0 \text{ and } (\nu_2 > 1 \text{ or } \nu_2 = 0)$$

must hold. This condition is always satisfied for priors 1 through 3. For the second case, the tail behavior is governed by

$$\delta^{-(k-1+2h+2\nu_2)/2}$$

and so the integral will exist if $k - 1 + 2h + 2\nu_2 > 2$. This reduces to $k > 3$ for prior 1, $k > 1$ for priors 2 and 3, and $k > 3 - 2\nu_2$ for prior 4. Keep in mind that *both* conditions need to be satisfied. Rather than repeat these arguments for the integrals that follow, the existence results are summarized in a table at the end of this section.

Returning to the estimation problem, the first quantity to compute is

$$\mathrm{E}(\delta|y) = \int\delta f(\delta)d\delta/g.$$

The next, and most useful quantity, is

$$\mathrm{E}(\theta_i|y) = \int[w_i\hat{\theta}_i + (1 - w_i)\hat{\mu}]f(\delta)d\delta/g.$$

The next quantity of interest is $\mathrm{Var}(\theta_i|y)$. From Section 2.7.2, two integrals are needed. The first one is similar to the one above. It is

$$\mathrm{E}(\theta_i^2|y) = \int[w_i\hat{\theta}_i + (1 - w_i)\hat{\mu}]^2 f(\delta)d\delta/g.$$

The second one is $\int v_{ii}p(\alpha,\delta|y)d\alpha d\delta$. With regard to α (see Section 3.4), v_{ii} contributes a multiplicative constant of α and so the integral with respect to α is similar to the one done in Section 3.6.

$$\begin{aligned}\int v_{ii}p(\alpha,\delta|\boldsymbol{y})d\alpha d\delta &= \int(1 - w_i)[1 + (1 - w_i)/w.]\delta\alpha p(\alpha,\delta|\boldsymbol{y})d\alpha d\delta \\ &= \int(1 - w_i)[1 + (1 - w_i)/w.]\delta \\ &\quad \times c(\delta)^{-(k-1)/2}h(\delta)[\Pi_i w_i/w.]^{1/2} \times \{\lambda_1 + \Sigma_{ij}P_{ij}(y_{ij} - \hat{\theta}_i)^2 \\ &\quad + [\lambda_2 + \Sigma_i(\hat{\theta}_i - \hat{\mu})^2 w_i]/\delta\}^{-(N+q-5)/2} \\ &\quad \times \Gamma[(N + q - 5)/2]2^{(N+q-5)/2}d\delta.\end{aligned}$$

$$\begin{aligned}\text{Let } f^*(\delta) = {} & (\delta)^{-(k-1)/2}h(\delta)[\Pi_i w_i/w.]^{1/2} \\ & \times \{\lambda_1 + \Sigma_{ij}P_{ij}(y_{ij} - \hat{\theta}_i)^2 \\ & + [\lambda_2 + \Sigma_i(\hat{\theta}_i - \hat{\mu})^2 w_i]/\delta\}^{-(N+q-5)/2} \\ & \div (N + q - 5)\end{aligned}$$

and so

$$\int v_{ii}p(\alpha,\delta|\boldsymbol{y})d\alpha d\delta = \int(1 - w_i)[1 + (1 - w_i)/w.]\delta f^*(\delta)d\delta/g.$$

In case there is interest in the two variance components, their posterior means are given by

$$E(\sigma^2|\boldsymbol{y}) = \int f^*(\delta)d\delta/g$$

and

$$E(\tau^2|\boldsymbol{y}) = \int \delta f^*(\delta)d\delta/g.$$

The predictive distribution for the next observation in the i^{th} class was discussed in Section 2.7.4. If the next X is $N(\theta_i,\sigma^2/R_i)$ then

$$E(X|\boldsymbol{y}) = E(\theta_i|\boldsymbol{y})$$

$$\text{Var}(X|\boldsymbol{y}) = E(\sigma^2|\boldsymbol{y})/R_i + \text{Var}(\theta_i|\boldsymbol{y}).$$

These quantities have already been obtained.

The final item to obtain by integration is the posterior density of θ_i. From Section 2.7.3,

$$\begin{aligned}p(\theta_i|\boldsymbol{y}) = {} & \int(s_i)^{-1}\delta^{-k/2}h(\delta)[\Pi_i w_i/w.]^{1/2} \\ & \times \{[(\theta_i - m_i)/\delta s_i]^2 + \Sigma P_{ij}(y_{ij})^2 - \Sigma_{ij}(\hat{\theta}_i)^2 P_i w_i \\ & - (\Sigma_i w_i\hat{\theta}_i)^2/w.\delta\}^{-(N+q-2)/2}d\delta \\ & \times \Gamma[(N + q - 2)/2]/\{\Gamma[(N + q - 3)/2]g\sqrt{\pi}\}\end{aligned}$$

where

$$m_i = w_i\hat{\theta}_i + (1 - w_i)\hat{\mu}$$

and

$$(s_i)^2 = (1 - w_i)[1 + (1 - w_i)/w.].$$

TABLE 1

CRITERIA FOR THE EXISTENCE OF THE INTEGRALS

	$\int f(\delta)d\delta$ $\int\tilde{\theta}_i f(\delta)d\delta$ $\int\tilde{\theta}_i^2 f(\delta)d\delta$	$\int\delta f(\delta)d\delta$	$\int f^*(\delta)d\delta$ $\int v_{ij}f^*(\delta)d\delta$	$\int\delta f^*(\delta)d\delta$
Prior 1	$k > 3$	$k > 5$	$k > 3$	$k > 5$
Priors 2,3	$k > 1$	$k > 3$	$k > 1$	$k > 3$
Prior 4	$k > 3 - 2\nu_2$	$k > 5 - 2\nu_2$	$k > 3 - 2\nu_2$	$k > 5 - 2\nu_2$
and $\lambda_2 > 0$	$N + 2\nu_1 > 3$	$N + 2\nu_1 > 3$	$N + 2\nu_1 > 5$	$N + 2\nu_1 > 5$
or $\lambda_2 = 0$	$\nu_2 > 1$ or $\nu_2 = 0$	$\nu_2 > 2$ or $\nu_2 = 1$	$\nu_2 > 1$ or $\nu_2 = 0$	$\nu_2 > 2$ or $\nu_2 = 1$

3.8 An EB Procedure Based on Integration

To do all the calculations listed in the previous section requires a large number of approximate integrations. It would be helpful if those that are needed for each of the k groups individually could be avoided. A compromise that is reminiscent of EB methodology is presented here.

To proceed it is necessary to obtain a general result for the mean and variance of a function of a random variable. To do this begin with a general random variable X and a function $g(x)$. Use the Taylor series expansion about $\xi = \mathrm{E}(X)$ to write

$$\mathrm{E}[g(X)] \doteq \mathrm{E}[g(\xi) + (X - \xi)g'(\xi)] = g(\xi)$$

and

$$\mathrm{Var}[g(X)] \doteq \mathrm{Var}[g(\xi) + (X - \xi)g'(\xi)] = [g'(\xi)]^2\mathrm{Var}(X).$$

Generally, for these approximations to be reasonable, the random variable X should in some sense be the average of a fairly large number of observations and the function $g(x)$ should be thrice differentiable around ξ. Almost any advanced text on mathematical statistics will contain theorems that make the above results precise. The random variable under consideration here is $\delta|y$ and its density will converge in the same manner in which a sample mean converges as the sample size, N, increases.

To evaluate $\mathrm{E}(\theta_i|y) = \mathrm{E}(w_i\hat{\theta}_i + (1 - w_i)\hat{\mu}|y)$, let $\delta|y$ play the role of X and so $\xi = \mathrm{E}(\delta|y) = \tilde{\delta}$, a quantity obtained in Section 3.7. Then let $\tilde{w}_i = P_i\tilde{\delta}/(1 + P_i\tilde{\delta})$. Finally,

$$\mathrm{E}(\theta_i|y) \doteq \tilde{w}_i\hat{\theta}_i + (1 - \tilde{w}_i)\tilde{\mu}$$

where $\tilde{\mu} = \Sigma\tilde{w}_i\hat{\theta}_i/\Sigma\tilde{w}_i$. The approximation, and the ones to follow, treated $\hat{\mu}$ as a fixed quantity to be replaced by the estimate obtained above. This seems reasonable as there should be little error involved in estimating the overall mean.

For the variance, begin with

$$\begin{aligned}\mathrm{Var}(\theta_i|y) &= \mathrm{Var}[w_i(\hat{\theta}_i - \hat{\mu}) + \hat{\mu}|y] + \mathrm{E}\{\alpha\delta(1 - w_i)[1 + (1 - w_i)/w.]|y\}\\ &\doteq (\hat{\theta}_i - \tilde{\mu})^2\mathrm{Var}(w_i|y) + \mathrm{E}\{\alpha\delta(1 - w_i)[1 + (1 - w_i)/w.]|y\}.\end{aligned}$$

For $\mathrm{Var}(w_i|y)$ use the approximation to obtain

$$\mathrm{Var}(w_i|y) \doteq (P_i)^2[\mathrm{E}(\delta^2|y) - \tilde{\delta}^2]/(1 + P_i\tilde{\delta})^4.$$

The second moment is obtained by numerically evaluating $\int\delta^2 f(\delta)d\delta/g$. For the expectation term, first integrate with respect to α. Following the development in Section 3.7 the expectation becomes

$$\mathrm{E}\{\delta(1 - w_i)[1 + (1 - w_i)/w.]|y\},$$

where the expectation is taken with respect to the density $f^*(\delta)/g$. Use the result on approximating expectations to get

$$\begin{aligned}\ddot{\delta} &= \int\delta f^*(\delta)d\delta/\int f^*(\delta)d\delta\\ &= \mathrm{E}(\tau^2|y)/\mathrm{E}(\sigma^2|y),\\ \ddot{w}_i &= P_i\ddot{\delta}/(1 + P_i\ddot{\delta})\end{aligned}$$

and the expectation term

$$\mathrm{E}(\sigma^2|y)\ddot{\delta}(1 - \ddot{w}_i)[1 + (1 - \ddot{w}_i)/\ddot{w}.].$$

Put all of the above together to obtain

$$\begin{aligned}\mathrm{Var}(\theta_i|y) \doteq\ &(\hat{\theta}_i - \tilde{\mu})^2(P_i)^2[\mathrm{E}(\delta^2|y) - \tilde{\delta}^2]/(1 + P_i\tilde{\delta})^4\\ &+ \mathrm{E}(\tau^2|y)(1 - \ddot{w}_i)[1 + (1 - \ddot{w}_i)/\ddot{w}.].\end{aligned}$$

4. A NUMERICAL ILLUSTRATION WITH THE ONE-WAY MODEL

In this section, two data sets are introduced to demonstrate the Bayesian analysis of the one-way model. The first set consists of workers' compensation frequency data. It will be used to provide some idea as to the computational resources needed. In addition, the influence of the three priors will be investigated as well as the quality of the EB style approximations. The second data set consists of workers' compensation loss ratios. This set will be used to demonstrate the evaluation of prediction intervals. In both cases, the results will be compared to those obtained from the usual EB formulas.

4.1 Workers' Compensation Frequency Data

The data were supplied by the National Council on Compensation Insurance (NCCI) and comprise observed frequencies from 7 years on 133 rating groups in 36 states. To make this data set somewhat manageable, the years were combined to yield the following:

y_{ij} = relative frequency in state j from group i

P_{ij} = Payroll in state j from group i.
$i = 1, \ldots, k = 133 \quad j = 1, \ldots, n_i$.

The number of states per group (n_i) is not always 36 as some states had no exposures for some groups. The total number of observations was $N = 4{,}572$, indicating that 216 cells had no exposure.

The objective is to estimate θ_i, the relative frequency of claims from insureds in rating group i. While the payrolls were adjusted for inflation, no attempt was made to adjust for any trend in the number of claims. As a final note, only claims resulting in permanent partial disability were included. It is important to recognize that the purpose of these illustrations is not to recommend a specific ratemaking procedure for workers' compensation insurance, but rather to illustrate the calculations using the formulas of Section 3. In particular, one might check the possibility that a cross-classified (state by group) model better describes the process.

For comparison, the formulas recommended by Bühlmann and Straub [7] were evaluated. These are the conventional EB formulas and produce the following results:

$$\hat{\sigma}^2 = \Sigma P_{ij}(y_{ij} - \hat{\theta}_i)^2/(N - k) = 4.746$$

$$\hat{\tau}^2 = [\Sigma P_i(\hat{\theta}_i - \hat{\mu})^2 - (k - 1)\hat{\sigma}^2]/(P - \Sigma P_i^2/P) = -.1123$$

where $\hat{\mu} = \Sigma P_i\hat{\theta}_i/P$ and $P = \Sigma P_i$.

When a negative value is obtained for τ^2, the convention is to use zero. This produces credibility weights of zero and so the grand mean is used as the estimate for each of the group means. As has been mentioned before, this method does not allow for evaluation of the quality of the estimates.

The key function in all the integrations is $f(\delta)$ as displayed in Section 3.7. The only item that involves the index j is $\Sigma(y_{ij} - \hat{\theta}_i)^2$ and it depends only on the data and so is constant. The number of observations per group will not affect the computation, other than the evaluation of a single sum of squares. All of the sums that involve δ have $k = 133$ terms. A second item is the size and shape of the integrand. The function $f(\delta)$ is well-behaved, being small at $\delta = 0$ and zero as $\delta \rightarrow \infty$, and having a single mode. However, because a variety of constants were removed from this density, it turned out that the value at the mode was very large. To see just how large, a variety of values of $\ln(f(\delta))$ were computed. In the actual calculations that followed, I worked with $f(\delta)/\exp(4{,}650)$ as the maximal value was near $\exp(4{,}650)$. Since all expressions are the ratio of two integrals involving this function, the adjustments cancel. Finally, it should be noted that $f(\delta)$ was calculated by first obtaining the logarithm of each of its constituent factors, adding them, subtracting 4,650, and then exponentiating the result. This avoided any overflow or underflow problems in the intermediate calculations.

In Table 2, the results for the first three priors are displayed. Note that while the posterior mean of τ^2 is indeed small (so zero was not an unreasonable estimate), it is large relative to σ^2/P_i, and so zero was not a reasonable choice for the credibility weight. This led to results that were considerably different from those obtained by the Bühlmann-Straub formula. Values of μ and the z_i were found by solving the following system of k equations:

$$E(\theta_i|y) \doteq z_i\hat{\theta}_i + (1 - z_i)\tilde{\mu} \text{ where } \tilde{\mu} = \Sigma z_i\hat{\theta}_i/\Sigma z_i.$$

The z_i then take on the role of credibility factors. These do not automatically arise in a Bayesian framework, and except for the fact that actuaries are accustomed to seeing this quantity, there is no reason to compute it. The three classes displayed were the ones with the smallest, median, and largest values of $\hat{\theta}_i$, respectively.

It is not surprising that the three priors produced virtually identical results. The large amount of data overwhelms all of these priors. In addition, I computed the standard deviation of $(\tau^2|y)$. Under prior 2 it is 0.000416. The mean of 0.003087 is over seven standard deviations above zero, indicating that the Bühlmann-Straub estimate is extremely unlikely to be valid.

The same items were evaluated using the approximations from Section 3.8. This was done only for prior 2, since the results will be similar for the others. The results are displayed in the last column of Table 2. In this case, the approximation performed well.

TABLE 2

Estimates by Integration in the One-Way Model

	Prior1	Prior2	Prior3	EBStyle
$E(\delta \| y)$	.0006613	.0006498	.0006503	$\tilde{\delta}$ .0006498
$E(\sigma^2 \| y)$	4.755	4.753	4.753	
$E(\tau^2 \| y)$	.003143	.003087	.003089	
Inferred μ	.07488	.07487	.07487	$\tilde{\mu}$ .07488
Class 107 (Auditors, Accountants, Draftsman)				
P_i	102,471	102,471	102,471	102,471
$\hat{\theta}_i$	.002762	.002762	.002762	.002762
z_i	.9852	.9849	.9850	$\tilde{w}_i$ .9852
$E(\theta_i \| y)$	.003830	.003848	.003847	.003829
$SD(\theta_i \| y)$	.006763	.006761	.006761	.006762
Class 68 (Explosives and Ammunition Mfg.)				
P_i	3,018	3,018	3,018	3,018
$\hat{\theta}_i$	.06262	.06262	.06262	.06262
z_i	.6638	.6599	.6600	$\tilde{w}_i$ .6623
$E(\theta_i \| y)$	.06674	.06679	.06678	.06676
$SD(\theta_i \| y)$	.03237	.03227	.03227	.03234
Class 89 (Stevedore)				
P_i	11,275	11,275	11,275	11,275
$\hat{\theta}_i$	.3895	.3895	.3895	.3895
z_i	.8800	.8782	.8783	$\tilde{w}_i$ .8799
$E(\theta_i \| y)$	.3518	.3512	.3512	.3517
$SD(\theta_i \| y)$	.01980	.01979	.01979	.01980
CPU (sec.)	14.92	15.23	15.11	12.49
Cost ($)	3.91	3.97	3.94	3.50

The computation was done on an IBM 4381 computer. The time did not include that used for setting up the data set (computing and arranging the values of y_{ij} and P_{ij}), so this can be viewed as the increase in cost of the Bayesian method over the Bühlmann-Straub formula (which is essentially free).

In addition, the iterative algorithm (Section 3.5) was employed with prior 2. Eight iterations were required for convergence. The results were $\tilde{\delta} = 0.0005936$, $\tilde{\sigma}^2 = 4.625$, and $\tilde{\mu} = 0.07479$. The results compare favorably with those obtained by integration.

4.2 Workers' Compensation Loss Ratio Data

This data set was taken from Meyers [25]. He provided loss ratios for three years of experience in 319 rating classes in the state of Michigan. In addition, the premium volume was given for each class/year; they will be used as the P_{ij} as in the Meyers paper. In that paper he used the Bühlmann-Straub formulas to obtain the credibility estimates. In view of the success from the previous section, I only computed estimates based on prior 2 and the EB approximation. The results were (the column labelled EB contains the results from the Meyers paper):

	EB	HNLM
σ^2	92,374	101,650
τ^2	0.019237	0.019762
μ	0.5822	0.5799
$K = \sigma^2/\tau^2$	4,801,900	5,143,710

It is not surprising that the results are similar. This also indicates that the Bühlmann-Straub formulas are indeed based on a hidden assumption of normality.

One of the most useful features of the Meyers data set is that it also provided the premiums and actual losses for the year following the three years of experience. This admits an evaluation of the predictive ability of the various procedures. I will begin the evaluation by duplicating the two tests performed by Meyers. In performing the tests, the expected losses based on the estimated loss ratios were adjusted to make the total expected losses equal to the actual losses. This is legitimate, since both credibility procedures were formulated to indicate relativities, not the absolute level of future losses. To do so would require trend factors to be incorporated into the analysis. An indication of how this might be done within the HNLM is given in the next section.

The first test is to measure the squared error of the predicted versus the observed losses. Bayes procedures (of any kind) should do well since the objective is to minimize squared error. The formula is $\Sigma P_i(A_i/E_i - 1)^2/k$ where A_i is the observed losses, E_i is the expected losses, and P_i is the premium. Also available for this test were the losses expected according to the rates promulgated by the NCCI. In addition, the weighted average relative error of the predictions was computed. The formula is $\Sigma P_i|A_i/E_i - 1|/\Sigma P_i$. The results were:

	Mean squared errors	Mean relative errors
NCCI	298,063	0.26776
EB	289,651	0.26396
HNLM	287,416	0.26368

The second test was invented by Meyers [25]. He called it the "Underwriting Test." The idea is to consider an insurer with established rates and a new entrant into the market. The new entrant uses his own method to determine premiums. He then offers insurance only to applicants in those rating classes for which his calculations produce rates less than those of the established insurer. He then charges a slightly lower premium than the established insurer and gets all of this business. If the new entrant's ratemaking methods are superior, he will expect a profit from his actions. Assuming differences only in relativities, but not in overall level, the established insurer will lose the same amount that the new entrant gains. A formalization of this process has A_i as the actual losses, E_i (for established) as the established insurer's expected losses, and N_i (for new) as the new entrant's expected losses. The profit and loss ratio, respectively, for the new entrant will be

$\Sigma(E_i - A_i)$ and $\Sigma A_i/\Sigma E_i$,

where all sums are taken over those classes for which $N_i < E_i$. The comparisons among the three estimators are presented in Table 3.

TABLE 3

THE UNDERWRITING TEST

Established	New Entrant	Profit for New Ent.	Loss Ratio
NCCI	EB	9,751,941	.950
NCCI	HNLM	9,929,786	.949
EB	NCCI	2,311,720	.990
EB	HNLM	7,493,090	.952
HNLM	NCCI	2,584,151	.989
HNLM	EB	−6,753,204	1.026

According to Meyers [25], a loss ratio of less than 0.957 has less than a five percent probability of occurring by chance. The results are consistent with the mean squared error ordering in that HNLM has a significant loss ratio as a new entrant against both EB and NCCI. Neither is significant as a new entrant against HNLM. By the same reasoning, EB is superior to NCCI. There is also an inconsistency present in that HNLM appears to do better versus NCCI than it does versus EB. An examination of the data reveals that the problem is the lack of sensitivity of this approach. Below are the results from two of the 319 classes (all figures in thousands of dollars):

		Expected		
Class	Loss	NCCI	EB	HNLM
8033	4,704	8,135	8,165	8,149
9079	22,208	15,464	16,087	16,108

In these classes HNLM scores a big "win" over EB, although the predictions are indistinguishable. Also note that in class 8033, HNLM defeats EB but loses to NCCI. It happened that there were no similar cases producing great gains under EB with a premium only slightly better than HNLM. The similarity of the expected losses should not have produced such a large overall difference between EB and HNLM. I attribute this to the method itself.

Recall that one of the stated advantages of HNLM is that it also produces prediction intervals for the future observations. This was done for the 319 classes. The standard deviations of the predictions were computed according to the approximation in Section 3.8. The formula is

$$\begin{aligned}\mathrm{Var}(\theta_i|y) \doteq{}& (\hat{\theta}_i - 0.5799)^2(P_i)^2(2.030 \times 10^{-15})/(1 + 1.957 \times 10^{-7}P_i)^4 \\ &+ (0.01993 + 3.842 \times 10^{-9}P_i)/(1 + 1.944 \times 10^{-7}P_i)^2.\end{aligned}$$

If all is well, the standardized actual losses (actual minus predicted divided by the standard deviation) should follow a normal distribution with mean zero and variance one. To see if that is so, two plots were prepared. Figure 1 shows a histogram of the 319 standardized losses. It is apparent that there is more skewness present than one would expect from a normal distribution. The chi-square goodness-of-fit test statistic using 20 intervals is 71.55. With 17 degrees of freedom there is clearly a lack of fit. Figure 2 is a plot of the standardized errors against the expected losses. This can be used to check for serial correlations and constant variances. The former is not a problem and that is confirmed by performing a sign test. There are 153 sign changes out of 318 opportunities, clearly close to the expected number of 159. There does appear to be a problem

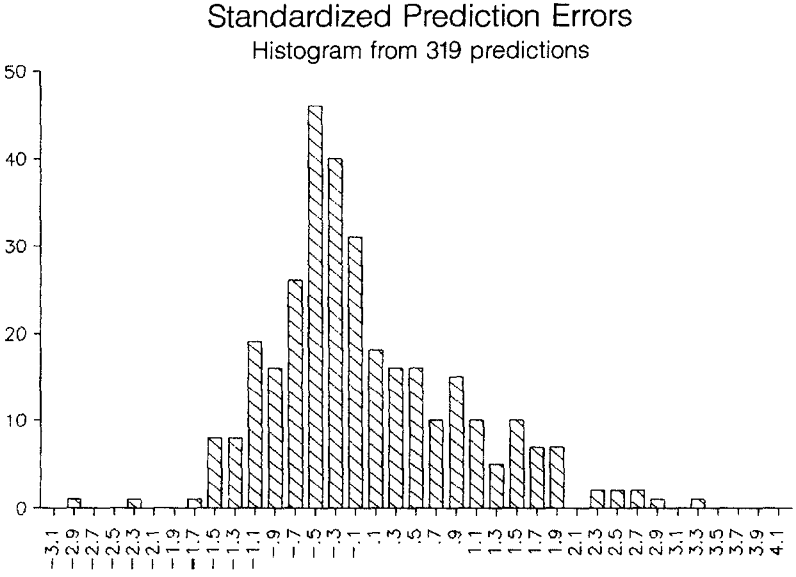

Figure 1

with the variances. For small predicted losses, the points are much too concentrated about the horizontal axis. This would lead us to suspect that we are overstating the variances in this range. One way to allow for this would be to adopt the unequal variance model of Section 5.1.

With these problems in mind, is there any value in attempting to predict these values? I believe there is. First of all, as was stated in Section 1, an inadequate model is almost always better than none at all. Secondly, we have some idea of the shortcomings, and could make some ad hoc corrections in the future.

As an illustration of the benefit of knowing the prediction errors, consider the following analysis which is done in the spirit of the "Underwriting Test." Identify all classes for which the HNLM predicted loss exceeds the NCCI predicted loss by at least *k* standard deviations. Do not offer insurance to these classes. For $k = 1$ there is only one class, number 4420. In thousands the

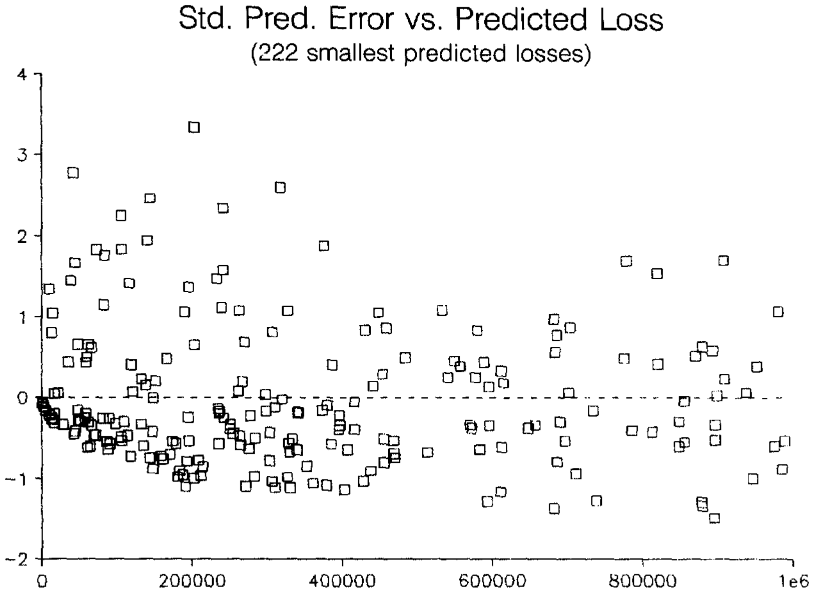

Figure 2

predictions were 4,329 and 3,427. The actual loss turned out to be 3,855. The point here is that this analysis can help identify classes in which the current rate levels are out of line, perhaps inspiring an investigation to see if something unusual has happened, either to the insureds in that class, or to the data in the process of recording it.

A final comment is in order. The above analysis leads us to believe that the normal model is not appropriate for these losses. In most settings we would not have the actual losses available in order to check this out. Can this be done with the original data? Box [4],[5] suggests the following approach to model-checking. In the general Bayesian setting, let X be the marginal distribution of the observations. Its density is computed from $f(x) = \int f(x|\theta)f(\theta)d\theta$ where, as usual, $f(x|\theta)$ is the model density and $f(\theta)$ is the prior density. If the model and prior are reasonable, the observed data x should, in some sense, be a "typical" observation from this density. While Box suggests a specific test, I will just display the standardized observations.

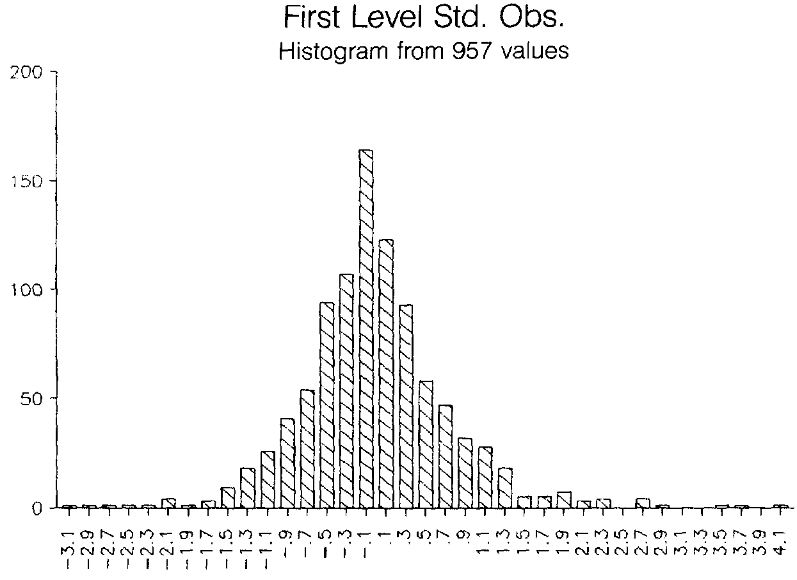

Figure 3

In particular, I will restrict attention only to the assumption of normality and will condition on the other aspects of the model such as constant variance. I will also condition on the estimated values of the variances. In the general HNLM, the distributions of interest are

First level—$\mathbf{y} \sim N(A\hat{\theta},\hat{G})$,

Second level—$\hat{\theta} \sim N(B\hat{\mu},\hat{H})$, and

Overall—$\mathbf{y} \sim N(\mathrm{AB}\hat{\mu},\hat{G} + A\hat{H}A')$.

In the particular case of the one-way model, these distributions become

First level—$y_{ij} \sim N(\hat{\theta}_i,\hat{\sigma}^2/P_{ij})$,

Second level—$\hat{\theta}_i \sim N(\hat{\mu},\hat{\tau}^2)$, and

Overall—$y_{ij} \sim N(\hat{\mu},\hat{\sigma}^2/P_{ij} + \hat{\tau}^2)$.

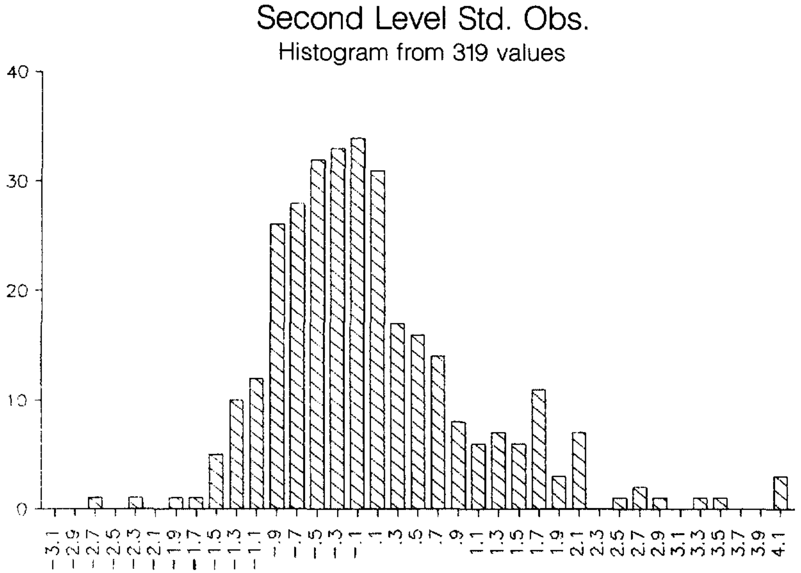

Figure 4

Figures 3, 4, and 5 display the histograms for the three sets of observations. In each case the appropriate values (y_{ij} or $\hat{\theta}_i$) were standardized according to the indicated means and variances. If the normal model was correct, the histograms should correspond to the standard normal distribution. An examination of the figures indicates that normality might indeed hold at the first level, but definitely does not at the second level. As a result, it is clear that the overall model should not be normal and that is indicated by the histogram.

Does the discovery of non-normality invalidate all the work that has been done? I believe the answer is no. We are at least as well off as one who used the EB methodology and we have the additional knowledge that we do not have the optimal solution. It is now a matter of deciding if the extra effort of analyzing a non-normal model is justified. Perhaps the ideas suggested in Section 1.3 are worth investigating.

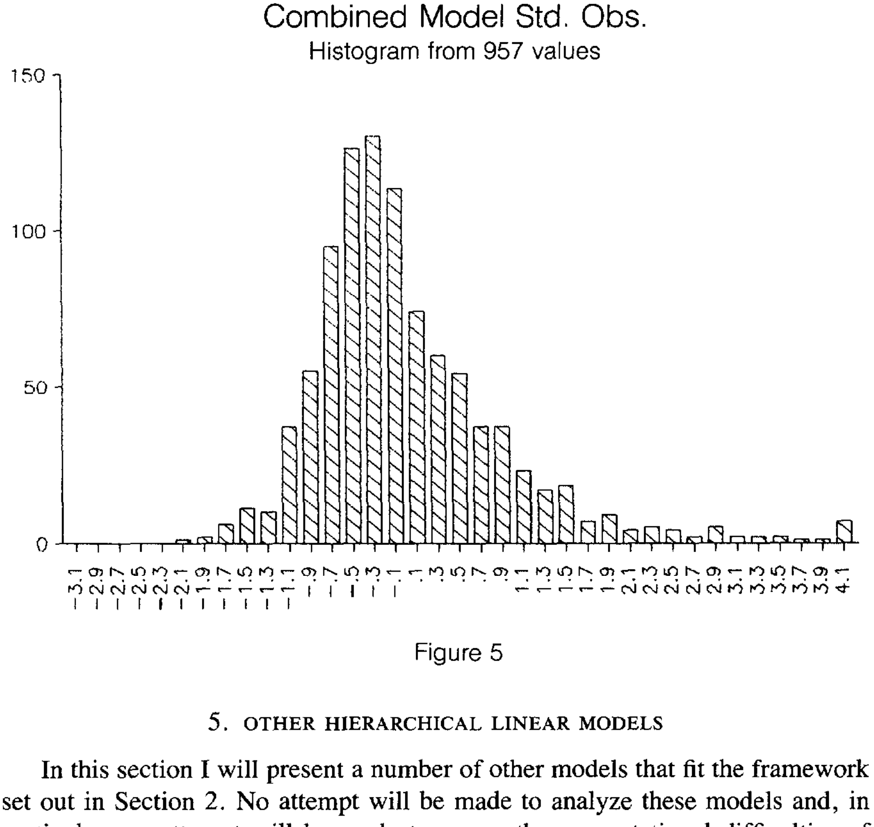

Figure 5

5. OTHER HIERARCHICAL LINEAR MODELS

In this section I will present a number of other models that fit the framework set out in Section 2. No attempt will be made to analyze these models and, in particular, no attempt will be made to assess the computational difficulties of evaluating these models. Unless there is an indication to the contrary it should be assumed that all of the random variables at a given level of the model are conditionally independent.

5.1 Unequal Variances

The process variance within each class may differ from class to class. From year to year within one class it is still assumed that variances are proportional to some exposure measure. The first two levels of the model are

Level 1—$Y_{ij}|\theta_i \sim N(\theta_i,\sigma_i^2/P_{ij})$ and

Level 2—$\theta_i|\mu \sim N(\mu,\tau^2)$.

Noninformative priors would then be placed on μ, τ^2, and $\sigma_1^2, \ldots, \sigma_k^2$.

It is easy to see what the EB approach to this model would be. Each σ_i^2 would be estimated from data in the i^{th} class. This is not in the spirit of credibility analysis where we would expect that information from the other classes can improve the estimation of a particular σ_i^2. A model that would do this would have an additional component at Level 2 such as

$$\sigma_i^2|\nu,\lambda \sim \text{Inverse gamma}(\nu,\lambda).$$

Noninformative priors would then be placed on μ, τ^2, ν, and λ.

5.2 *Parameter Uncertainty*

Suppose it is possible that the class mean θ_i varies from year to year, but not in any predictable manner. A model for this would be

Level 1—$Y_{ij}|\alpha_{ij} \sim N(\alpha_{ij},\sigma^2/P_{ij})$

Level 2—$\alpha_{ij}|\theta_i \sim N(\theta_i,\gamma^2)$, and

Level 3—$\theta_i|\mu \sim N(\mu,\tau^2)$.

Collapse the first two levels to produce

$$Y_{ij}|\theta_i \sim N(\theta_i,\sigma^2/P_{ij} + \gamma^2)$$

This is similar to a model proposed in Meyers [26]. It is not possible to derive EB estimates of the three variance terms as the within sum of squares is all that is available to estimate both σ^2 and γ^2. There is, however, a least squares approach based on the relationship of the variance in one group to its exposure that can yield estimates of the three parameters. A detailed HNLM analysis of this model is presented in Klugman [23]. The major problem is the evaluation of a two dimensional integral.

5.3 *Hierarchical*

In this paper, the word hierarchical applies to all the models. In credibility work this term has been reserved for the case where the k classes can be divided into g groups, where the i^{th} group would have m_i classes in it ($m_1 + \ldots + m_g = k$). Begin with a three level model:

Level 1—$Y_{ijt}|\theta_{ij} \sim N(\theta_{ij},\sigma^2/P_{ijt})$

Level 2—$\theta_{ij}|\beta_i \sim N(\beta_i,\gamma^2)$

Level 3—$\beta_i|\mu \sim N(\mu,\tau^2)$

Noninformative priors would be required for μ, τ^2, γ^2, and σ^2. Levels 2 and 3 may be combined to form a single distribution. However, when conditioned only on μ, the θ_{ij} are no longer independent. EB formulas for this model and the one in Section 5.4 are given in Venter [34].

5.4 Cross-Classified

Suppose each rating class is identified by two variables, such as sex and age, or state and occupation. An additive model, with the possibility of error, can be expressed with three levels:

Level 1—$Y_{ijt}|\theta_{ij} \sim N(\theta_{ij},\sigma^2/P_{ijt})$

Level 2—$\theta_{ij}|\mu,\alpha_i,\beta_j \sim N(\mu + \alpha_i + \beta_j,\gamma^2)$

Level 3—$\alpha_i \sim N(0,\tau_1^2)$
$\beta_j \sim N(0,\tau_2^2)$

Noninformative priors would be placed on μ, τ_1^2, τ_2^2, γ^2, and σ^2. The first two levels are easily collapsed to produce the single level

$$Y_{ijt}|\mu,\alpha_i,\beta_j \sim N(\mu + \alpha_i + \beta_j,\gamma^2 + \sigma^2/P_{ijt}).$$

It has been common to set $\gamma^2 = 0$ in analyzing this model. Including it allows for some departure from additivity. Letting τ_1^2 and τ_2^2 become infinite (so uniform priors are placed on all α_i and β_j) produces a simple version of the model. The credibility compromise is between a strict additive model and the use of individual class means.

5.5 Linear Trend

In the one-way model we might observe that there is a year to year trend in the means. A simple linear trend would be modeled as

Level 1—$Y_{ij}|\alpha_i,\beta_i \sim N(\alpha_i + j\beta_i,\sigma^2/P_{ij})$

Level 2—$\alpha_i|\mu_1 \sim N(\mu_1,\tau_1^2)$
$\beta_i|\mu_2 \sim N(\mu_2,\tau_2^2)$

with noninformative priors on μ_1, μ_2, τ_1^2, τ_2^2, and σ^2. This is similar to the well-known model introduced by Hachemeister [13]. It could be generalized to other types of trend by altering the structure of the mean at level 1.

5.6 Time Series

A linear time series model can be formulated in two stages. Here the subscripts indicate observations at a given time t.

Level 1—$\mathbf{Y}_t|\boldsymbol{\theta}_t \sim N(F_t\boldsymbol{\theta}_t, A)$

Level 2—$\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1} \sim N(G_t\boldsymbol{\theta}_{t-1},B)$

The matrices F_t and G_t are known while A and B require prior distributions. The first level is the process distribution which explains how the observations relate to the underlying parameters. The second level is the state distribution which explains how the parameters change over time.

As an example, consider the linear trend model from Section 5.4. In this setting it would look like

Level 1—$Y_{ij}|\theta_{ij} \sim N(\theta_{ij},\sigma^2/P_{ij})$

Level 2—$\theta_{ij}|\theta_{i,j-1} \sim N(\beta_i + \theta_{i,j-1},\tau^2)$.

It is not exactly the same, as level 2 implies that there are some disturbances that let the progression of means depart from strict linearity. The parameter α_i in Section 5.5 is unchanging over time. Prior distributions would be needed for $\theta_{i,0}$ (to get the system started) and for β_i, τ^2, and σ^2.

This model is very similar to the Kalman filter. An excellent non-Bayesian application of this model to loss reserving is found in deJong and Zehnwirth [18]. A discussion of its relationship to the usual credibility models is given in deJong and Zehnwirth [19].

6. CONCLUSIONS AND AREAS FOR FUTURE RESEARCH

The intent of this paper was to introduce the hierarchical normal linear model as a tool for classification ratemaking. This model has three advantages over the EB approach. First, methods for estimating the variances do not have to be created on a case-by-case basis. Instead, the estimates fall naturally out of the analysis. Second, estimates of estimation and prediction error are available. Finally, model-checking and model-selection procedures can be employed. The latter was not discussed in this paper, but methods do exist for identifying the most appropriate model when there are several to choose from (for example, a one-way vs. a cross-classified analysis). See Klugman [23] for an application.

Of course, this approach also introduces difficulties of its own. Foremost among them are the intensive computations needed to perform the analysis. In addition, the derivation of formulas for specific models can be very time consuming (although once obtained they can be used over and over). These problems are really another advantage of the HNLM approach; they are all technical in nature and are certain to be solved if there is sufficient interest in doing so.

The major area for future work (other than grinding out the solution to the many models of interest) is the relaxation of the normality assumption. There is overwhelming evidence that insurance data are not normal and so methods to accommodate that fact are most desirable. I envision two ways to attack this problem. One is to create methods that are robust against general departures from normality. To do this, the *t*-distribution or a mixture of normal distributions could be used in the model. Another way would be to find methods that are superior under specific distribution assumptions that are likely to correspond to insurance experience. In any case, considerable sensitivity testing should be done to any recommended formula.

REFERENCES

[1] R. Bailey, "A Generalized Theory of Credibility," *PCAS* XXXII, 1945, p. 13.

[2] J. Berger, *Statistical Decision Theory and Bayesian Analysis*, Second ed., New York, NY, Springer-Verlag, 1985.

[3] G. Box and G. Tiao, *Bayesian Inference in Statistical Analysis*, Reading, MA, Addison-Wesley, 1973.

[4] G. Box, "Sampling and Bayes' Inference in Scientific Modelling and Robustness," *Journal of the Royal Statistical Society*, Series A, Vol. 143, 1980, p. 383.

[5] G. Box, "An Apology for Ecumenism in Statistics," *Proceedings of the Conference on Scientific Inference, Data Analysis, and Robustness*, G. Box, T. Leonard, and C-F. Wu, eds., New York, NY, Academic Press, 1983.

[6] H. Bühlmann, "Experience Rating and Credibility," *ASTIN Bulletin*, Vol. 4, 1967, p. 199.

[7] H. Bühlmann and E. Straub, "Credibility for Loss Ratios," *ARCH*, 1972.2.

[8] R. Burden, J. Faires, and A. Reynolds, *Numerical Analysis*, Third ed., Boston, MA, PWS, 1981.

[9] W. Ericson, "A Note on the Posterior Mean of a Population Mean," *Journal of the Royal Statistical Society*, Series B, Vol. 31, 1969, p. 332.

[10] S. Geisser, "On Prior Distributions for Binomial Trials (with discussion)," *The American Statistician*, Vol. 38, 1984, p. 244.

[11] P. Goel, "On Implications of Credible Means Being Exact Bayesian," *Scandinavian Actuarial Journal*, 1982, p. 41.

[12] F. Graybill, *Matrices with Applications in Statistics*, Second ed., Belmont, CA, Wadsworth, 1983.

[13] C. Hachemeister, "Credibility for Regression Models with Applications to Trend," *Credibility: Theory and Applications*, P. Kahn, ed., New York, NY, Academic Press, 1975.

[14] C. Hewitt, "Credibility for Severity," *PCAS* LVII, 1970, p. 148.

[15] R. Hogg and S. Klugman, *Loss Distributions*, New York, NY, Wiley, 1984.

[16] Insurance Services Office, *Report of the Credibility Subcommittee: Development and Testing of Empirical Bayes Procedures for Classification Ratemaking,* New York, NY, ISO, 1980.

[17] W. Jewell, "Credible Means are Exact Bayesian for Exponential Families," *ASTIN Bulletin,* Vol. 8, 1974, p. 77.

[18] P. deJong and B. Zehnwirth, "Claims Reserving, State-Space Models and the Kalman Filter," *Journal of the Institute of Actuaries,* Vol. 110, 1983, p. 157.

[19] P. deJong and B. Zehnwirth, "Credibility Theory and the Kalman Filter," *Insurance: Mathematics and Economics,* Vol. 2, 1983, p. 281.

[20] S. Klugman, "Loss Distributions: Estimation, Large Sample Theory, and Applications," *Premium Calculation in Insurance,* F. de Vylder, M. Goovaerts, and J. Haezendonck, eds., Dordrecht, Holland, Reidel, 1984, p. 263.

[21] S. Klugman, "Distributional Aspects and Evaluation of Some Variance Estimators in Credibility Models," *ARCH,* 1985.1, p. 73.

[22] S. Klugman, "Bayesian Credibility with a Noninformative Prior," *Insurance and Risk Theory,* M. Goovaerts, F. de Vylder, and J. Haezendonck, eds., Dordrecht, Holland, Reidel, 1986, p. 195.

[23] S. Klugman, "Credibility Under Parameter Uncertainty: An Illustration of the Hierarchical Normal Linear Model," paper submitted for publication, 1987.

[24] D. Lindley and A. Smith, "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society,* Series B, Vol. 34, 1972, p. 1.

[25] G. Meyers, "Empirical Bayesian Credibility for Workers' Compensation Classification Ratemaking," *PCAS* LXXI, 1984, p. 96.

[26] G. Meyers, "An Analysis of Experience Rating," *ARCH,* 1985.2, p. 373.

[27] C. Morris, "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association,* Vol. 78, 1983, p. 47.

[28] C. Morris, "Parametric Empirical Bayes Confidence Intervals," *Proceedings of the Conference on Scientific Inference, Data Analysis, and Robustness,* G. Box, T. Leonard, and C-F. Wu, eds., New York, NY, Academic Press, 1983.

[29] J. Nelder and R. Mead, "A Simplex Method for Function Minimization," *Computer Journal,* Vol. 6, 1965, p. 308.

[30] R. Norberg, "Empirical Bayes Credibility," *Scandinavian Actuarial Journal,* 1980, p. 177.

[31] M. Novick and P. Jackson, *Statistical Methods for Educational and Psychological Research,* New York, NY, McGraw-Hill, 1974.

[32] A. Smith, A. Skene, J. Shaw, J. Naylor and M. Dransfield, "The Implementation of the Bayesian Paradigm," *Communications in Statistics, Theory and Methods,* Vol. 14, 1985, p. 1079.

[33] G. Tiao and A. Zellner, "Bayesian Estimation of Multivariate Regression," *Journal of the Royal Statistical Society,* Series B, Vol. 26, 1964, p. 277.

[34] G. Venter, "Structured Credibility in Applications—Hierarchical, Multidimensional, and Multivariate Models," *ARCH,* 1985.2, p. 267.

[35] A. Whitney, "The Theory of Experience Rating," *PCAS* IV 1918, p. 274.