Generalized Linear Mixed Models for Ratemaking: A Means of Introducing Credibility into a Generalized Linear Model Setting

Fred Klinker, FCAS, MAAA

Abstract: GLMs that include explanatory classification variables with sparsely populated levels assign large standard errors to these levels but do not otherwise shrink estimates toward the mean in response to low credibility. Accordingly, actuaries have attempted to superimpose credibility on a GLM setting, but the resulting methods do not appear to have caught on. The Generalized Linear Mixed Model (GLMM) is yet another way of introducing credibility-like shrinkage toward the mean in a GLM setting. Recently available statistical software, such as SAS PROC GLIMMIX, renders these models more readily accessible to actuaries. This paper offers background on GLMMs and presents a case study displaying shrinkage towards the mean very similar to Buhlmann-Straub credibility.

Keywords: Credibility, Generalized Linear Models (GLMs), Linear Mixed Effects (LME) models, Generalized Linear Mixed Models (GLMMs).

1. INTRODUCTION

Generalized Linear Models (GLMs) are by now well accepted in the actuarial toolkit, but they have at least one glaring shortcoming--there is no statistically straightforward, consistent way of incorporating actuarial credibility into a GLM.

Explanatory variables in GLMs can be either continuous or classification. Classification variables are variables such as state, territory within state, class group, class within class group, vehicle use, etc. that take on a finite number of discrete values, commonly referred to in statistical terminology as "levels." The GLM determines a separate regression coefficient for each level of a classification variable. To the extent that some levels of some classification variables are only sparsely populated, there is not much data on which to base the estimate of the regression coefficient for that level. The GLM will still provide an estimated coefficient for that level but will assign it a large standard error of estimation. In effect, the GLM warns the user to exercise considerable care in interpreting that coefficient but doesn't otherwise adjust the estimated coefficient to take into account the low volume of data. When faced with this situation, the natural inclination of an actuary is to shrink low credibility levels towards the mean, but the GLM quotes a large standard error of estimation and leaves it at that.

There have been a number of responses to this unsatisfactory state of affairs. Some actuaries have been known to apply an ad hoc credibility adjustment to coefficients output by a GLM. In some cases this even produces results similar to those arrived at by more statistically rigorous methods. If so, then what is so wrong with the ad hoc credibility adjustment of GLM output? First, we cannot guarantee the ad hoc results will always agree closely with results from those other methods. Second, the statisticians who designed our GLMs were unaware we intended to subject GLM estimates to the violence of a subsequent round of ad hoc credibility adjustments. If they had known, they might have suggested a better starting point than GLM estimates. This gets back to the old issue that a sequence of steps, each optimal individually, may not be optimal in the aggregate.

Turning to other, more statistically rigorous attempts to incorporate credibility in a GLM setting, it would be desirable to find a method that estimates both the GLM and the credibility adjustment in a single, statistically consistent step where each GLM estimation and credibility adjustment takes into account the fact that the other estimation process is also going on. A number of authors have indeed produced models that combine GLM and credibility, for example, Nelder and Verrall (1997), Ohlsson and Johansson (2004 and 2006), and Ohlsson (2006). Given the importance of the issue these papers address, why have these models not caught on in actuarial circles (at least not that I am aware)? I might hazard two guesses. First, their math is somewhat complex and perhaps intimidating. Second, their algorithms are iterative and require a nontrivial degree of programming from their users.

There are alternative statistical models, quite similar in theory to Nelder and Verrall and Ohlsson and Johansson, known as Generalized Linear Mixed Models (GLMMs). Statisticians actually developed these models some time ago, but it has only been very recently that popular stat software (like SAS, R, and S-Plus) has been enhanced to provide us with the means to readily estimate these models. Furthermore, it should be noted that models much like GLMMs have even been introduced into the actuarial literature. See, for example, Guszcza (2008), which admittedly introduced these models in a reserving rather than ratemaking setting, but that paper does provide a good introduction and intuition regarding what is going on in the guts of a GLMM, or something much like a GLMM.

I will not argue that GLMMs provide better models than Nelder and Verrall and Ohlsson and Johansson, but the newly available software makes it easier to implement these GLMM models.

1.1 Objectives of This Paper

The objectives of this paper are to:

- Introduce Linear Mixed Effects (LME) models and their generalization to Generalized Linear Mixed Models (GLMMs).

- Show how the LME, at least when applied to simple models, can be solved in closed form and leads directly to shrinkage of random effects towards the mean of the form of Buhlmann-Straub credibility. This motivates at least the hope that a similar shrinkage might be expected from a GLMM, where the math is no longer tractable in closed form.
- Demonstrate the application of a GLMM to a case study in which the hoped-for shrinkage is indeed observed and does indeed approximate the form of Buhlmann-Straub credibility.
- Demonstrate along the way (in Appendix A) SAS code that implements the GLMM.

In one sense, the central point of the paper is Table 4 and Figure 1. These show the shrinkage observed in the case study and the fact that this shrinkage is of form approximately Buhlmann-Straub. The reader who takes nothing else away from this paper should at least keep this Table and Figure in mind as motivation for wanting to learn more about GLMMs as a means of implementing credibility in a GLM setting in a manner very reminiscent of credibility theory they already know.

The reader should also keep in mind what this paper is *not*, and the following comments are offered as a means to managing readers' expectations. What this paper is not is a general review article on the various means by which credibility has been incorporated into a GLM setting. I will not discuss the various ad hoc credibility adjustments to GLM output alluded to earlier, nor Nelder and Verrall, nor Ohlsson and Johansson, nor other more overtly Bayesian or Empirical Bayes methods. I will not discuss the subtle theoretical points in which Nelder and Verrall and Ohlsson and Johansson differ from GLMMs nor examine the differences in results produced by applying ad hoc credibility adjustments vs. Nelder and Verrall vs. Ohlsson and Johansson vs. GLMMs to the case study of this paper. What this paper is intended to say is, "Here is one very interesting way of implementing credibility in a GLM setting. It might or might not be the best from among those methods currently available, but it is certainly promising. It produces credibility-like shrinkage very similar to credibility you, the actuary, are already familiar with. And it has the added advantage of ready implementation via software only recently available."

This paper is also not intended to be a comparison of GLMM implementations in different stat packages. SAS PROC GLIMMIX is likely to be available to many of the readers of this paper, and it happens to be the means I chose to implement GLMMs. But there are also implementations in R, S-Plus, etc., and I don't mean to imply that SAS PROC GLIMMIX is superior to these others.

I leave to another, more energetic and ambitious author the task of writing the general review article that some readers might have been hoping for. Indeed, I would hope that this paper serve as the impetus for such a review article.

1.2 Prerequisites

It will be assumed that the reader is already familiar with the theory of GLMs and their application to actuarial problems at the level of McCullagh and Nelder (1989), Anderson et al. (2004), and de Jong and Heller (2008).

1.3 Outline of Remainder of This Paper

The remainder of this paper proceeds as follows. Section 2 will introduce the Linear Mixed Effects (LME) model, a simpler cousin of the GLMM, as a means of introducing many features of GLMMs before I discuss their complications. Section 3 will show how, in a very simple case, Buhlmann-Straub credibility emerges directly from the LME model. This is done to motivate the connection between GLMMs and credibility. The LME is generalized to the GLMM in Section 4.

Section 5 presents a case study on live ISO data for an unspecified line of business. By comparing GLM and GLMM runs on essentially the same model form and same data, it is shown that the GLMM introduces a shrinkage of sparsely populated classification variable levels towards the mean. This shrinkage is not seen in the GLM. Furthermore, it is shown that the credibility implied by this shrinkage is very close to the form of Buhlmann-Straub credibility. Section 6 concludes. SAS code implementing the Section 5 case study as well as a discussion of some of the output from that code has been deferred to Appendix A.

2. THE LINEAR MIXED EFFECTS (LME) MODEL

The Linear Mixed Effects model is nothing other than classical linear regression (more correctly, the classical general linear model) with the addition of "random" effects to the "fixed" effects already treated in classical linear regression. The error distribution is assumed normal. Expected values are assumed linear in explanatory variables. In the language of GLMs, the error distribution in the "exponential" family is the normal distribution, and the link function is the identity function.

Especially because it is so central to the understanding of the rest of this paper, more needs to be said about the distinction between "fixed" and "random" effects. The classic fixed effect is a classification variable with relatively few levels, those being the only levels we are interested in. The classic random effect is a classification variable with potentially many levels, only some of which appear in our dataset by design of the sample that produced our dataset from the overall population. Re random effects, the focus is frequently on the variance among the levels rather than on the values of the levels themselves, which are assumed to have expectation zero. Even when there is interest in the values of the levels of the random effects, the inferential algorithm that predicts those levels must first estimate the random effects variances. It should also be noted that to some extent the distinction between fixed and random effects depends on the context of the study; the very same effect treated as fixed in one study might reasonably be treated as random in another, given the different goals of the two studies.

Consider the following example. Suppose you want to test the relative efficacy of a number of drugs, so your model includes a drug main effect. The levels of that drug main effect test whether some drugs are better than others, better than a control, better than a placebo, etc. in terms of some response that serves as the dependent variable in your model. You run your drug trials at a number of test centers. This suggests that you include test center as another main effect in your model to control for possible test center differences. If you treat test center as a fixed effect, you end up drawing inferences for drug main effects appropriate for those test centers but not validly extendible to medical centers other than those at which you ran the tests. On the other hand, if you drew your test centers relatively at random from a much larger universe of possible centers, and you reflect that fact by treating test center as a random main effect in your model, then you end up drawing drug main effect inferences that can validly be extended to centers other than the ones at which you actually did the tests. Quoted standard errors of the drug effects will be somewhat larger because of the additional uncertainty attributable to the treatment center random effect. For further enrichment re random effects, see Littell et al. (2006) or any one of many texts on random or mixed effects models.

The above describes the classical statistical motivation for random effects, but the actuarial motivation for considering random effects actually differs somewhat from this. If we treat state, or territory within state, as random, is that because we want to extend our model results to states or territories we didn't actually have in our data? No, not usually. If we want our model results to apply to a given state or territory, we usually include that state or territory in our data. But the zero expectation of random effects models creates a natural drift towards zero when data is thin. This shrinkage towards zero looks a lot like credibility with the complement of credibility being the overall mean, and we invoke the statistical machinery of random effects to exploit that shrinkage.

Turning to the description of the LME model, the basic equation is:

$$Y = X\beta + Zu + e. \tag{2.1}$$

This would be the classic linear regression equation (expressed in matrix form) except for the additional random effects term Zu. This is a matrix equation. If there are n observations, Y is an n-vector of the observed values of the dependent variable. X is an n by p matrix, referred to as the model structure (or design) matrix, for the fixed effects. If the model includes an intercept term, there is a column of X consisting of all ones to capture the intercept. For each continuous explanatory variable, there is a column of X consisting of the values for that variable. For each main effect classification variable with m levels, there are m columns of X which are indicator variables for membership in each of the m levels (except that certain full rank parameterizations of the model suppress one of the columns). The indicator variable for the i^{th} level takes the value one if the observation is indeed in that level, zero otherwise. Interaction terms contribute more complex columns to X. β is the p-vector of regression coefficients for the fixed effects. This is not to say that there are a total of p fixed effect variables in the model, only that, taking into account the intercept and the fact that classification variables contribute multiple columns to the structure matrix, it requires a total of p regression coefficients to fully specify the fixed effects part of the model.

Z is the n by q design matrix for the model random effects. The columns of Z are indicator variables for membership in classification variable levels for those classification variables treated as random effects. The q-vector u is the equivalent of β and can be thought of as the vector of regression coefficients for the random effects. The n-vector e is the vector of random measurement errors.

Further structure is imposed by the following assumptions. Both u and e vectors are multivariate normally distributed with expectations 0. The variance of u is a q by q matrix Var[u]=G. The variance of e is an n by n matrix Var[e]=R. The u and e vectors are assumed uncorrelated: Cov[u,e]=0. The structure of G specifies the structure of correlation among the random effects. G is frequently assumed diagonal or block diagonal. Other types of correlational structure among the observations, such as autocorrelated time series or spatial structure, are specified through the structure of the R matrix. The user will most likely specify the structure of G and R, but these matrices may include unknown parameters that have to be estimated as part of the LME algorithm. It is common to speak of G side and R side covariance structure to distinguish between correlation arising through random effects vs. other time series or spatial processes. These G side and R side covariance structures tend to be specified in different places in the model specification syntax.

This model structure gives rise to two relevant distributions. (The reader is forewarned that this dichotomy will become more important when the LME is generalized to the GLMM in a later section.) The first is the conditional distribution, $Y|u$, of the dependent variable Y conditional on actually knowing the random effects u. This distribution has expectation $X\beta+Zu$ and variance R. The second distribution is the marginal distribution for Y not knowing the random effects, which is the conditional distribution integrated over the random effects. It has expectation $X\beta$ (because u has expectation 0) and variance $V=\mathrm{Var}[Zu+e]=ZGZ'+R$, where Z' denotes the transpose of the Z matrix. Note that the total variance V has G side and R side contributions. In a normal world, where both u and e are multivariate normal, so are both the conditional and marginal distributions for Y, but this result need not extend to the GLMM.

The LME is solvable in closed form via generalized least squares. The estimator for β is BLUE, Best Linear Unbiased Estimator (or EBLUE, Estimated or Empirical Best Linear Unbiased Estimator, if the total variance matrix, V, includes unknown parameters that have to be estimated as part of the solution. For a discussion of BLUE and related terms, see Littell et al. (2006)), and is given by

$$\hat{\beta} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}Y. \tag{2.2}$$

The predictor for the random effects is the expectation of the random effects u conditional on the observed Y, is BLUP, Best Linear Unbiased Predictor (or EBLUP, Estimated or Empirical Best Linear Unbiased Predictor), and is given by

$$E[u \mid Y] = E[u] + \mathrm{Cov}[u,Y]\mathrm{Var}[Y]^{-1}(Y - E[Y]) = GZ'V^{-1}\left(Y - X\hat{\beta}\right). \tag{2.3}$$

Discussion of how the unknown parameters in V are estimated would take us too far afield, nor is it necessary for the following argument.

The reader seeking further enrichment re LME models can consult such books as Littell et al. (2006). This book also provides discussion of such standard statistical terminology as BLUE and BLUP, the distinction between estimators for fixed effects and predictors for random effects, the generalization from LME to GLMM, as well as numerous examples of implementations of LMEs and GLMMs via SAS software.

3. HOW BUHLMANN-STRAUB CREDIBILITY EMERGES FROM THE LME MODEL

A key point of this paper is that random effects in LME and their generalization to GLMMs entail a credibility-like shrinkage. This result is exact for LME and particularly easy to see in the following simple example.

Let us assume that class is the only explanatory variable, so we are looking at a one-way ANOVA, treating the grand mean as a fixed effect and the class offsets about the grand mean as random effects. We assume data has been aggregated, so there is only one observation per class, the class i average response y_i. Y is the vector of the y_i. Exposures for class i are w_i. X is the fixed effects design matrix appropriate for an intercept only model, hence a single column of identical ones. β, the vector of fixed effects regression coefficients, is only a 1-vector with the single entry the intercept. The random effects design matrix, Z, has columns that are indicator variables for the various levels of the class variable. If we assume our observations are in class order (first observation in the first class, second observation in the second class, etc.), then Z is just an identity matrix.

We turn next to the structure of the R side and G side variance matrices. R is the variance of the random errors e. We assume the e are independent of one another from class to class, so R is diagonal. Furthermore, we assume the i^{th} class error variance is equal to a proportionality constant known as the within variance, σ_w^2, divided by the exposure volume w_i. In other words, the error variance declines with increasing volume. So R is diagonal with diagonal elements σ_w^2/w_i. The random effects u are also assumed independent from class to class, so G is also diagonal with diagonal elements equal to the so-called between variance σ_b^2. Then the total variance matrix $V=ZGZ'+R$ is also diagonal with diagonal elements

$$V_i = \sigma_b^2 + \frac{\sigma_w^2}{w_i}. \tag{3.1}$$

We will not here address the estimation of the unknown within and between variances but treat them for present purposes as known. In fact, with only one observation per class, the within variance may not even be estimable. The reader should also note that by reference to "within" and "between variance" we have slipped into actuarial jargon; to my knowledge "within" and "between variance" are not common statistical terms.

The estimator for the grand mean becomes, with very little algebra, exploiting the many structural simplifications of this particular example,

$$\hat{\beta} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}Y = \left(\sum_i \frac{1}{\sigma_b^2 + \frac{\sigma_w^2}{w_i}}\right)^{-1} \sum_i \frac{y_i}{\sigma_b^2 + \frac{\sigma_w^2}{w_i}}. \tag{3.2}$$

Defining credibility as

$$z_i = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma_w^2}{w_i}} = \frac{w_i}{w_i + \frac{\sigma_w^2}{\sigma_b^2}} \quad , \tag{3.3}$$

equation (3.2) becomes

$$\hat{\beta} = {\sum_i z_i y_i} \Big/ {\sum_i z_i} \, . \tag{3.4}$$

So the BLUE estimator of the fixed effects grand mean is none other than the credibility weighted average of the class means.

Turning next to the prediction of the random class effects, we already know by assumption that the unconditional expectations of the random effects vanishes, E[u]=0, and we know the total variance Var[Y]=V is diagonal. We can also show in the present case that the covariance matrix Cov[u,Y] is diagonal.

$$\begin{aligned} \text{Cov}[u_i, y_i] &= \text{Cov}[u_i, \beta + u_i + e_i] = \text{Var}[u_i] = \sigma_b^2. \\ \text{Cov}[u_i, y_j] &= \text{Cov}[u_i, \beta + u_j + e_j] = \text{Cov}[u_i, u_j] = 0. \end{aligned} \tag{3.5}$$

Collecting these results into the generic BLUP predictor of equation (2.3), the diagonality of the matrices on the right-hand side of (2.3) causes the matrix equation to collapse to a collection of scalar equations in which u_i depends only on y_i.

$$E[u_i \mid Y] = E[u_i \mid y_i] = \text{Cov}[u_i, y_i]\text{Var}[y_i]^{-1}\left(y_i - \hat{\beta}\right)$$
$$= \frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma_w^2}{w_i}}\left(y_i - \hat{\beta}\right)$$
$$= z_i\left(y_i - \hat{\beta}\right). \quad (3.6)$$

So the posterior predictor of the random effect represents a shrinkage of the observed class mean towards the fixed effects grand mean by a factor that amounts to Buhlmann-Straub credibility. (For the reader needing a refresher on Buhlmann-Straub credibility, see chapter 4 of Goovaerts and Hoogstad (1987).) This might give us reason to hope that a similar result might hold at least approximately when LME is generalized to GLMM.

4. GENERALIZATION OF LME TO THE GENERALIZED LINEAR MIXED MODEL (GLMM)

For actuarial applications, the most restrictive assumptions in the LME model are that errors are normally distributed and that expected values are linear in explanatory variables. We already know how much power is gained by generalizing the classical linear model to the GLM and would hope for a similar gain in power on applying similar generalizations to the LME model.

- In the notation of the previous section, the conditional distribution Y|u is now assumed to be in the exponential family rather than normal. Recall that the normal is a special case of the exponential family.
- Rather than assuming the conditional expectation linear in explanatory variables, we assume there is at least a link function g such that the g transform of the conditional expectation is linear: g(E[Y|u])=Xβ+Zu. Note that the identity link is a special case of this assumption.
- We still assume u multivariate normal with mean 0, variance matrix G, and uncorrelated with the random measurement error Y-E[Y|u].

The resulting model is the Generalized Linear Mixed Model (GLMM). Because the normal distribution is a special case of the exponential family, and the identity link is a special case of a more general link function, the LME model is a special case of the more general GLMM.

However, there are a number of important features of the LME model that do not carry over to the GLMM. One of them has to do with marginal distributions. In LME, the conditional distribution Y|u is normally distributed. So is u. The marginal distribution Y, being the integration of Y|u over u, is also normal. This does not always extend to the GLMM. Y|u is distributed in the exponential family. The random effect u is still assumed normally distributed. But the marginal distribution Y may not even be in the exponential family. Keep that in mind as you interpret GLMM output.

A second complication is that GLMM equations are not usually solvable in closed form. Instead, there are iterative solution algorithms, much as for the GLM. As a further consequence, there is no closed form algebra producing Buhlmann-Straub credibilities as we observed above in the LME case. But we can compare a GLM and a GLMM run on the same data and essentially the same model form to find evidence of shrinkage in the GLMM not present in the GLM. By plotting this shrinkage against measures of volume, we find evidence that the shrinkage is fit closely by credibility of Buhlmann-Straub form. This is demonstrated in the case study of the next section of this paper.

5. A CASE STUDY

5.1 Structure of the Problem

This case study is based on live, not simulated, ISO data. I have masked both line of business as well as names of potential explanatory variables to preserve ISO's intellectual property. The dependent variable being modeled is experience ratio, the ratio of observed losses to expected losses under the current rating plan, the latter denoted as ALCCL. The data are not at the level of individual risks but rather aggregated into cells defined as crossings on all relevant explanatory variables, producing about 300,000 cells. As a consequence, some cells contain only a few risks; others contain thousands. One might therefore expect choice of proper weights to be important. Classical actuarial reasoning would lead us to expect, just as in the case of loss ratios, that the variance of experience ratios be inversely proportional to volume. In other words, the weights should be some measure of volume of business. I tested a number of possible weights and observed the best weight diagnostics when I used ALCCL as weights.

Before delving further into detail, we might fruitfully give some brief thought to what it means to model on experience ratios as the dependent variable. The denominator of the experience ratio is in effect the current rating plan. If the current rating plan is entirely adequate, we would expect to see no statistically significant evidence of structure in the experience ratio across the explanatory variable space, so any statistically significant evidence of structure is evidence for changes to the current rating plan, and the model parameters indicate the degree of change.

I limited the case study model to four explanatory variables so as not to swamp the case study with too much detail. The variable of primary interest is a classification variable, CLASS1, with twelve levels. Some of these levels are sparsely populated, and we will therefore want to treat CLASS1 as a random effect in a GLMM so as to shrink those levels towards the grand mean. There are two other classification variables, COVARIATE1 and COVARIATE2, each with four levels, which we will treat as fixed effects. There is also a continuous variable, COVARIATE3.

By default, SAS encodes classification variable effects in linear models (and their generalizations, such as GLM and GLMM) as contrasts between each level and the last level in the list for that variable. So, for each of CLASS1, COVARIATE1, and COVARIATE2, I have selected a well-populated level to serve as the base level for that variable and recoded it to "9" or "99" to force it to the end of the list of levels for that variable. This means that all contrasts will be expressed relative to stable bases.

Note that we are not modeling frequency and severity separately but rather their joint impact on experience ratio. We therefore need a distribution with positive mass at zero (to capture those cells with no loss) as well as a continuous density on the positive reals to capture cells with loss. It has recently become popular to model such cases with a Tweedie distribution with exponent p between 1 and 2. An exponent p=1.67 is a popular choice, and that is what I have chosen for the present case study.

Finally, we assume the ever-popular natural log link so as to yield a multiplicative model.

5.2 The GLM

The GLM model as summarized above was estimated via SAS PROC GENMOD. Further detail re code, etc. is deferred to Appendix A. The resulting parameter estimates from that model are summarized in the following Table 1.

Table 1

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.2963	0.0329	-0.3608	-0.2317	80.96	<.0001
class1	01	1	0.3346	0.0413	0.2538	0.4155	65.78	<.0001
class1	02	1	0.2585	0.1674	-0.0696	0.5865	2.38	0.1225
class1	03	1	0.3056	0.0463	0.2149	0.3963	43.59	<.0001
class1	04	1	-0.1181	0.0774	-0.2697	0.0335	2.33	0.1267
class1	05	1	0.4388	0.1278	0.1882	0.6894	11.78	0.0006
class1	06	1	0.2196	0.0487	0.1242	0.3149	20.36	<.0001
class1	07	1	0.4695	0.0885	0.2960	0.6430	28.12	<.0001
class1	08	1	0.4268	0.0928	0.2449	0.6086	21.15	<.0001
class1	10	1	0.2978	0.0708	0.1591	0.4366	17.71	<.0001
class1	11	1	-0.1779	0.1198	-0.4126	0.0569	2.20	0.1376
class1	13	1	-0.0423	0.1865	-0.4077	0.3232	0.05	0.8207
class1	99	0	0.0000	0.0000	0.0000	0.0000	.	.
covariate1	2	1	0.0704	0.0367	-0.0014	0.1423	3.69	0.0548
covariate1	3	1	-0.0507	0.0840	-0.2152	0.1139	0.36	0.5460
covariate1	4	1	-0.3958	0.1005	-0.5927	-0.1989	15.52	<.0001
covariate1	9	0	0.0000	0.0000	0.0000	0.0000	.	.
covariate2	3	1	-0.1554	0.0531	-0.2595	-0.0514	8.57	0.0034
covariate2	4	1	-0.0778	0.0406	-0.1574	0.0019	3.66	0.0557
covariate2	5	1	0.0617	0.1423	-0.2172	0.3407	0.19	0.6646
covariate2	9	0	0.0000	0.0000	0.0000	0.0000	.	.
covariate3		1	-0.3652	0.0697	-0.5019	-0.2285	27.42	<.0001
Scale		0	670.2088	0.0000	670.2088	670.2088		

Note first that CLASS1 level 99, COVARIATE1 level 9, and COVARIATE2 level 9 all have 0 degrees of freedom (DF) and 0.0000 estimates and standard errors of those estimates, because these are the base levels for their respective classification variables, are therefore pegged at 0.0000, and all other levels are expressed as contrasts off these. Second, some of the CLASS1 levels have much larger standard errors than others. These are the poorly populated levels most in need of credibility treatment. More will be said about the Scale parameter in Appendix A.

5.3 The GLMM

Now it is desired to give CLASS1 a credibility treatment, so the GLMM model as summarized above was estimated via SAS PROC GLIMMIX, treating CLASS1 as a random effect. Again, further detail re code, etc. is deferred to Appendix A, but considerably more detail is provided for the GLMM relative to the GLM, given that the primary focus of this paper is on GLMMs.

Table 2 displays the resulting fixed effects parameter estimates and Table 3 the random effects parameter estimates, specifically for the CLASS1 variable.

Table 2

Solutions for Fixed Effects							
Effect	**covariate1**	**covariate2**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept			-0.08888	0.06507	11	-1.37	0.1993
covariate1	2		0.07203	0.03666	312E3	1.96	0.0494
covariate1	3		-0.04689	0.08411	312E3	-0.56	0.5772
covariate1	4		-0.3965	0.1006	312E3	-3.94	<.0001
covariate1	9		0	.	.	.	.
covariate2		3	-0.1547	0.05310	312E3	-2.91	0.0036
covariate2		4	-0.07826	0.04070	312E3	-1.92	0.0545
covariate2		5	0.05959	0.1421	312E3	0.42	0.6749
covariate2		9	0	.	.	.	.
covariate3			-0.3643	0.06961	312E3	-5.23	<.0001

These estimates are quite similar to those from the GLM (compare Tables 1 and 2) with the exception of the intercept. This is because CLASS1 is treated as a fixed effect in the GLM, centered about its level 99, and is treated as a random effect in the GLMM, centered about a mean value of approximately 0. The different centering of CLASS1 between GLM and GLMM results in offsetting adjustments to the intercepts in the two models. Standard errors of the fixed effects are also quite similar between the two models.

Table 3

Solution for Random Effects						
Effect	**class1**	**Estimate**	**Std Err Pred**	**DF**	**t Value**	**Pr > |t|**
class1	01	0.1241	0.06917	312E3	1.79	0.0729
class1	02	0.03040	0.1334	312E3	0.23	0.8198
class1	03	0.09508	0.07121	312E3	1.34	0.1818
class1	04	-0.2898	0.08770	312E3	-3.30	0.0010
class1	05	0.1674	0.1161	312E3	1.44	0.1492
class1	06	0.01150	0.07222	312E3	0.16	0.8735
class1	07	0.2229	0.09499	312E3	2.35	0.0190
class1	08	0.1836	0.09723	312E3	1.89	0.0589
class1	10	0.08142	0.08468	312E3	0.96	0.3363
class1	11	-0.2876	0.1106	312E3	-2.60	0.0093
class1	13	-0.1349	0.1391	312E3	-0.97	0.3322
class1	99	-0.2040	0.06847	312E3	-2.98	0.0029

Note that there is no preferred base level; a parameter is quoted for every level, and the parameters appear to be approximately mean zero.

5.4 Inferred Credibility

We now extract the CLASS1 credibilities implicit in the GLMM by comparing the CLASS1 parameter output from the GLMM to that from the GLM. We do this in the following Table 4.

Table 4

Inferred Credibility

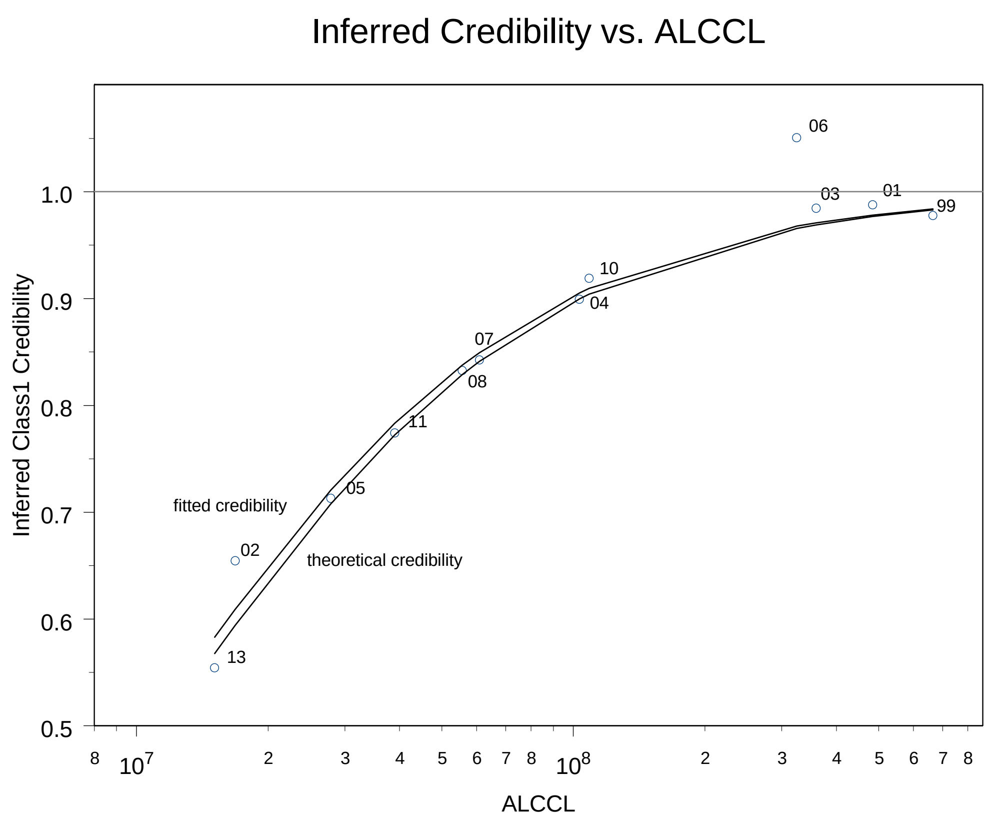
		Fixed Effects			Random Effects			(9)
		(3)		(5)	(6)		(8)	Inferred
(1)	(2)	Class1	(4)	Class1	Class1	(7)	Class1	Class1
Class1	ALCCL	Effect	exp(effect)	Relativity	Effect	exp(effect)	Relativity	Credibility
01	484,185,185	0.3346	1.3974	1.1417	0.1241	1.1321	1.1399	0.9877
02	16,832,999	0.2585	1.2950	1.0580	0.0304	1.0309	1.0380	0.6545
03	359,748,011	0.3056	1.3574	1.1090	0.0951	1.0998	1.1073	0.9845
04	103,293,336	-0.1181	0.8886	0.7260	-0.2898	0.7484	0.7536	0.8994
05	27,864,645	0.4388	1.5508	1.2671	0.1674	1.1822	1.1904	0.7129
06	324,592,379	0.2196	1.2456	1.0176	0.0115	1.0116	1.0185	1.0506
07	60,941,612	0.4695	1.5992	1.3066	0.2229	1.2497	1.2583	0.8426
08	55,682,170	0.4268	1.5323	1.2519	0.1836	1.2015	1.2098	0.8328
10	108,633,028	0.2978	1.3469	1.1004	0.0814	1.0848	1.0923	0.9190
11	39,019,053	-0.1779	0.8370	0.6839	-0.2876	0.7501	0.7552	0.7742
13	15,101,361	-0.0423	0.9586	0.7832	-0.1349	0.8738	0.8798	0.5542
99	664,914,612	0.0000	1.0000	0.8170	-0.2040	0.8155	0.8211	0.9777
	2,260,808,391		1.2240	1.0000		0.9932	1.0000	

Fixed effect parameter estimates (from GLM) are tabulated in column (3), random effect parameter estimates (from GLMM) in column (6). But these parameters reside in the space of the linear predictor. To put them in the scale of the original observations, we invert the log link in columns (4) and (7). As already noted, the fixed effect parameters are expressed as contrasts to level 99, which was chosen for its volume, and hence stability, rather than for its being relatively centered among the levels. So we would not expect the mean of fixed parameters to be near 0, nor the mean of their exponentials to be near 1, and indeed they are not. Dividing the column (4) exponentials by their mean (weighted on ALCCL) produces relativities relative to a mean relativity of 1 in column (5). Due to the manner in which they were predicted, the random effect parameters are far closer to mean 0, but we still adjust column (7) to a mean relativity of 1 in column (8).

The column (5) and column (8) relativities are now directly comparable. If credibility is implicit in a GLMM, the column (8) random effect relativities should have shrunk towards 1 relative to the column (5) fixed effect relativities. Defining inferred credibility as column (9) = (column (8) - 1)/(column (5) - 1), the evidence is there. Furthermore, plotting these inferred credibilities against

ALCCL reveals evidence of declining credibility with declining volume. (See Figure 1. The two curves will be discussed in the next subsection of this paper.)

Figure 1



One anomaly clearly stands out, CLASS1 level 06, for which the credibility is considerably in excess of 1.000. If we examine Table 4, we find that, for level 06, both fixed and random effect relativities are so close to 1.000 that even small errors or distortions in those relativities are magnified in the ratio that defines the inferred credibility. Among possible sources of error in the credibilities could be the fact that the renormalization of relativities to a mean of 1.0000 from column (3) to (5) and from column (6) to (8) in Table 4 introduces correlations among the CLASS1 level parameters not present in the output from the GLM and GLMM.

5.5 Inferred Credibility Is Approximately in Buhlmann-Straub Form

Are the above credibilities in approximately Buhlmann-Straub form

$$c = \frac{w}{w+k} \tag{5.1}$$

where c is credibility, w is a volume measure, and k is a constant? The trick is to determine k. If we assume Buhlmann-Straub form, then we can rework equation (5.1) into the following form:

$$\frac{1}{c} - 1 = \frac{k}{w} \tag{5.2}$$

This suggests we define a dependent variable equal to the reciprocal of our inferred credibilities minus 1 and regress this against an explanatory variable equal to reciprocal ALCCL in a regression through the origin (no intercept). The resulting regression coefficient would be our desired k. Applying this program to our Table 4 results, we find a k of 10.8 million (dollars). This regression is a simple, unweighted one. One could perhaps argue whether a weighted regression would be more appropriate, but this first approximation should suffice.

One can alternatively estimate k from certain parameters in the GLMM output. Appendix B derives a value of 11.5 million, in close agreement with the 10.8 million from the above regression.

Returning to expression (5.1) we substitute ALCCL for w and the two estimates for k, and plot the two resulting curves on Figure 1. The two curves are very close to each other and fit the inferred credibilities quite nicely. One may conclude that the implicit GLMM credibilities, at least for this case, are close to Buhlmann-Straub form.

6. SUMMARY AND CONCLUSIONS

GLMs signal, by quoting large standard errors, uncertain estimates for sparsely populated levels of classification explanatory variables, but they do not also adjust those estimates closer to the mean in response to low credibility. As a consequence, actuaries have desired for some time to introduce credibility into a GLM setting. There have been various attempts, both ad hoc and statistically rigorous, but none appear to have become popular, for reasons not always obvious.

The Generalized Linear Mixed Model (GLMM) provides yet another means of introducing credibility-like shrinkage into a GLM setting. Recently available statistical software, including SAS PROC GLIMMIX as well as new R and S-Plus functions, brings these models within reach of actuaries.

This paper first introduced the reader to the Linear Mixed Effects (LME) model, a simpler cousin of the GLMM, as a means of introducing issues important for GLMMs but in a less complex environment. It was shown how Buhlmann-Straub credibility falls directly out of the LME math, at least for a simple case. The LME was then generalized to the GLMM, and a case study demonstrated how to use GLMM software and showed that the GLMM preserved shrinkage to the mean in a form at least approximating Buhlmann-Straub credibility.

It is hoped that this paper will give actuaries sufficient knowledge, incentive, and courage to experiment with GLMMs in their next GLM project. New software, such as SAS PROC GLIMMIX, provides them the means to do this.

Acknowledgment

The author wishes to thank a number of reviewers for their helpful suggestions, especially those that resulted in tightening the argument and focusing in on the most important points of the paper.

Appendix A: SAS Implementation of the GLM and GLMM of the Case Study: Additional Detail

I have included this appendix for those readers who would like more detail on how to implement the GLMM of the case study in at least one stat package. This is not to imply that the SAS implementation of GLMMs is better than others, only that SAS is the package I chose. There are implementations of GLMM in R and S-Plus as well as other stat packages.

The reader should recall the model basics. The dependent variable is EXPRATIO, or experience ratio, assumed Tweedie distributed with exponent p equal to 1.67. Explanatory variables are classification variables CLASS1, COVARIATE1, and COVARIATE2, as well as the continuous variable COVARIATE3. CLASS1 is treated as a fixed effect in the GLM and a random effect in the GLMM. All other explanatory variables are treated as fixed effects in both the GLM and the GLMM. The regressions are weighted on ALCCL, a measure of business volume in each record. The link is log.

A.1 The GLM

The SAS code for the GLM is as follows:

```
PROC GENMOD DATA=INDATA;
 P=1.67;
 Y=_RESP_;
 A=_MEAN_;
 VARIANCE BAR=A**P;
 DEVIANCE DEV=2*((Y**(2-P)-Y*A**(1-P))/(1-P)-(Y**(2-P)-A**(2-P))/(2-P));
CLASS CLASS1 COVARIATE1 COVARIATE2;
WEIGHT ALCCL;
MODEL EXPRATIO= CLASS1 COVARIATE1 COVARIATE2 COVARIATE3/
      LINK=LOG SCALE=PEARSON;
RUN;
```

SAS PROC GENMOD does not naturally support the Tweedie distribution, but it does support a facility to allow users to specify their own distributions (by specifying both a variance and a deviance function for their distribution of choice). Lines 2 through 7 of the above code are what specify the Tweedie. The SCALE=PEARSON option in the MODEL statement is also important. The variance law for the Tweedie (which specifies the functional form of the observation variances) is $Var[y]=\varphi\mu^p/w$, where y is the observation, μ is the fitted value, p is the Tweedie exponent, w is the weight, and φ is the so-called dispersion coefficient. We are here telling the GLM to use a Pearson chi-squared estimator of the dispersion coefficient rather than assuming it equal to 1. The dispersion coefficient is fundamental, because it is the basis for the standard error estimates of the GLM coefficients. The Scale parameter of Table 1 of this paper is the square root of the estimated dispersion coefficient, and, at 670, is certainly quite far from 1.

A.2 The GLMM

Now we want to give CLASS1 a credibility treatment. The following code fits a GLMM to the same data to which we previously fit a GLM, and with as much of the same model form as before as possible, with the exception that CLASS1 is now treated as a random effect.

```
PROC GLIMMIX MAXOPT=50 PCONV=.000015 DATA=INDATA;
 _VARIANCE_=_mu_**1.67;
CLASS CLASS1 COVARIATE1 COVARIATE2;
WEIGHT ALCCL;
MODEL EXPRATIO= COVARIATE1 COVARIATE2 COVARIATE3/ LINK=LOG SOLUTION;
RANDOM CLASS1/ SOLUTION;
RANDOM _RESIDUAL_;
RUN;
```

Because a secondary purpose of this paper is to convince the reader that GLMMs are now within reach of actuaries via currently available software, we will spend more time on this code and its resulting output than we did on the prior GLM. First, technically, GLIMMIX doesn't fit a maximum likelihood but rather a maximum pseudo-likelihood. This means that, although you still need to specify the variance law of the Tweedie distribution (see line 2 of the code), you do not also need to specify a deviance function. Had you been interested in one of the distributions supported by GLIMMIX rather than the user-defined Tweedie, just as in GENMOD you would have specified that distribution via a DIST= option in the MODEL statement.

The MODEL statement specifies the fixed effects part of the model (and an option to the MODEL statement specifies the optional R side of the variance model). The random effects, which determine the G side of the variance model, are specified by the RANDOM statements. PROC GENMOD automatically gives you tables of parameter estimates and their standard errors, but, if you want those from PROC GLIMMIX as well, you have to ask for them via the SOLUTION options in the MODEL statement (for the fixed effects regression parameters) and in the RANDOM statement (for random effects regression parameters).

The RANDOM _RESIDUAL_ statement is crucial. GLIMMIX estimates a dispersion coefficient only for non-user-defined distributions, and even then only for those with a dispersion coefficient in their definition; otherwise, GLIMMIX pegs the dispersion coefficient at 1.000 by default. The RANDOM _RESIDUAL_ statement is the way in which you force GLIMMIX to still estimate the dispersion coefficient for user-defined distributions.

Lastly, note the MAXOPT and PCONV options in the PROC GLIMMIX statement. By default, GLIMMIX attempts 20 iterations of a certain outer iteration (the fact that there is also an inner iteration will be noted momentarily) before giving up. Furthermore, it determines model convergence when the percentage change of certain parameters from one iteration to the next is less than about 10^{-8}. The case study data was sufficiently volatile that the algorithm was never able to attain this high standard, but, by examining iteration history details, it was found that convergence could be achieved with a slightly relaxed standard of 50 iterations specified by the MAXOPT option and a convergence criteria of 1.5×10^{-5} specified by the PCONV option.

Before discussing GLIMMIX output, I will sketch an outline of what GLIMMIX is actually doing when it estimates the model, as this will aid interpretation of subsequent output. Unlike other familiar SAS model-building PROCs, GLIMMIX does not build models on the original data but rather on pseudo-data. At the beginning of each iteration, it constructs new pseudo-data by linearizing the original data about the expected values from the prior iteration. It then maximizes the pseudo-likelihood on that pseudo-data. It should also be noted that the iteration referred to here is the outer iteration.

The algorithm doesn't even solve simultaneously for the variance components (the unknown parameters of the total variance matrix V) and the fixed and random effects parameters. Rather, it starts the iteration with a pseudo-likelihood that is a function of variance components, fixed effects, and the dispersion coefficient and is able to adjust out ("profile" out) the fixed effects and dispersion coefficient to produce an objective function that is a function of just the variance components. It then enters an inner iteration to optimize this modified objective function over just the variance components. Armed with estimates of the variance components from the inner iteration it then estimates fixed effects and predicts random effects, then returns to the outer iteration for another pass through, starting with producing the next pseudo-data set, and so on until convergence. Although we will not examine an iteration history table output by our GLIMMIX run, if you were to examine such a table, you would note reference to iterations, restarts, and subiterations, which hints at the structure of inner iterations (subiterations) nested within outer iterations mentioned above.

Turning to the output of the above SAS PROC GLIMMIX code, following a first table that summarizes the dataset, the dependent variable, the assumed distribution, the link function, the weights, and a few other model assumptions, there are tables of additional model dimensions, shown as Tables 5 and 6, that are highly useful for checking that the model estimated is the one the user intended to estimate, and that there hasn't been some misinterpretation through some syntax error.

Table 5

Dimensions	
G-side Cov. Parameters	1
R-side Cov. Parameters	1
Columns in X	10
Columns in Z	12
Subjects (Blocks in V)	1
Max Obs per Subject	312131

Table 5 is crucial for verifying that GLIMMIX correctly interpreted our model specification. We did indeed want one G side parameter to be estimated, the between variance for the CLASS1 random effect. We did indeed want one R side variance parameter, the dispersion coefficient. There should indeed be ten columns in the fixed effects design matrix, one for the intercept term, four for each of COVARIATES 1 and 2, and one for COVARIATE 3. There should indeed be twelve columns in the random effects design matrix because CLASS1 has twelve levels.

Table 6

Optimization Information	
Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	1
Lower Boundaries	1
Upper Boundaries	0
Fixed Effects	Profiled
Residual Variance	Profiled
Starting From	Data

Table 6 reminds us that the inner iteration does indeed profile out both fixed effects and the dispersion coefficient (Residual Variance), and optimization is only over the remaining G side and R side variance components, in this case, over the single unknown parameter of the between variance of the CLASS1 random effect. Hence the optimization is only over one parameter. Because this parameter represents a variance, it is bounded below by zero, hence the reference to one lower boundary. But it is unbounded above.

Table 7 displays the resulting estimated variance component parameters.

Table 7

Covariance Parameter Estimates		
Cov Parm	**Estimate**	**Standard Error**
class1	0.04050	0.02095
Residual (VC)	449268	1137.28

As already noted, only two parameters were requested, the between variance of the CLASS1 random effect and the dispersion coefficient (Residual). The dispersion coefficient is in close agreement with that from the previous GLM. (Recall that the GLM dispersion coefficient is the square of the Scale parameter in Table 1: $670^2 = 448{,}900$.) The CLASS1 between variance is about .04. Its square root of about .2 (interpretable as approximately 20% because of the log link and because the random effects reside in the space of the linear predictor, in other words, in the logged space) indicates that levels of the CLASS1 variable fall a few tens of percent above and below the grand mean. This seems reasonable.

Following these tables presented above the SAS output provides fixed effects and random effects parameters already presented and discussed as Tables 2 and 3 of this paper. The reader seeking further detail is referred to the SAS PROC GLIMMIX online manual.

Appendix B: Inferring a Buhlmann-Straub k Parameter from GLIMMIX Output

Buhlmann-Straub credibility is of the form of equation (5.1). The k parameter in that equation is frequently written as a ratio of within variance to between variance, which is how it appears in equation (3.3). Can we read from our GLIMMIX output the numbers we would need to estimate within and between variance, and hence the k parameter? Yes, but the reader is forewarned that the following is not a strict derivation but rather a plausibility argument. It should be enough to support the approximate magnitude of the k parameter but not its precise value.

First, if one studies derivations of Buhlmann-Straub credibility (see chapter 4 of Goovaerts and Hoogstad (1987)), one finds that what is referred to as the within variance is actually the proportionality constant in the relationship: observation variance proportional to reciprocal weights. Recall the Tweedie variance law: $\mathrm{Var}[y]=\varphi\mu^p/w$, where φ is the dispersion coefficient, μ the expected value of y, p the Tweedie exponent, and w the weight. Strictly speaking, the numerator of this law is not a constant because μ is not, being a function of explanatory variables. Nevertheless, it might be reasonable to equate the within variance in the Buhlmann-Straub k to $\varphi\langle\mu\rangle^p$, where $\langle\mu\rangle$ is mean expectation.

Next, the variance component indicated by GLIMMIX for the classification variable in question is almost the desired between variance, except that it is measured in the space of the linear predictor, where the random effects live, and not in the original scale of the observations, as needed for the Buhlmann-Straub k. We just have to back-transform via the inverse link function, invoking the following approximation: $\mathrm{Var}[g(x)]\approx g'(E[x])^2\mathrm{Var}[x]$. In other words, the variance of the g transform of x is approximately the variance of x times the square of the derivative of g evaluated at mean x. Here, x is the linear predictor, Var[x] is the variance component from GLIMMIX, g is the inverse log link, in other words, the exponential. Its first derivative is also the exponential. E[x] is the mean linear predictor. The exponential (the inverse link) of the mean linear predictor is approximately $\langle\mu\rangle$. Then the between variance we seek is approximately $\langle\mu\rangle^2$ times the variance component.

Now, we assumed the Tweedie exponent p to be 1.67. The GLIMMIX output tells us the dispersion coefficient φ is 449,000 and the CLASS1 variance component is .0405. GLIMMIX doesn't tell us, but we know from other checks of our dataset that $\langle\mu\rangle$ is approximately .9, in which case,

$$k = \frac{\text{within variance}}{\text{between variance}} \approx \frac{\varphi\langle\mu\rangle^{p}}{\langle\mu\rangle^{2}\,\text{variance component}}$$

$$= \frac{\varphi}{\langle\mu\rangle^{2-p}\,\text{variance component}} \approx \frac{449{,}000}{(.9)^{2-1.67}(.0405)} \approx 11.5\text{million.}$$

7. REFERENCES

[1] Anderson, Duncan, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, and Neeza Thandi, "A Practitioner's Guide to Generalized Linear Models," *CAS Discussion Paper Program*, 2004, pp. 1-116.

[2] de Jong, Piet, and Gillian Heller, *Generalized Linear Models for Insurance Data*, Cambridge, U.K.: Cambridge University Press, 2008.

[3] Goovaerts, M. J., and W. J. Hoogstad, *Credibility Theory*, Survey of Actuarial Studies No. 4, Nationale-Nederlanden N. V., 1987.

[4] Guszcza, James C., "Hierarchical Growth Curve Models for Loss Reserving," *Casualty Actuarial Society E-Forum*, Fall 2008, pp. 146-173.

[5] Littell, Ramon C., George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, and Oliver Schabenberger, *SAS for Mixed Models*, 2nd ed., Cary, NC: SAS Institute, 2006.

[6] McCullagh, P., and J. A. Nelder, *Generalized Linear Models*, 2nd ed., Chapman & Hall/CRC, 1989.

[7] Nelder, J. A., and Richard Verrall, "Credibility Theory and Generalized Linear Models," *ASTIN Bulletin* 27(1), 1997, pp. 71-82.

[8] Ohlsson, E., and B. Johansson, "Combining Credibility and GLM for Rating of Multi-Level Factors," *CAS Discussion Paper Program*, 2004, pp. 315-326.

[9] Ohlsson, E., and B. Johansson, "Exact Credibility and Tweedie Models," *ASTIN Bulletin* 36(1), 2006, pp. 121-133.

[10] Ohlsson, E., "Credibility Estimators in Multiplicative Models," Research Report 2006:3, Mathematical Statistics Stockholm University.

Author's Biography

Fred Klinker (FCAS, MAAA) is an actuarial manager with ISO's Modeling Division. He is involved in a number of collaborative projects with ISO's other major predictive modeling unit, ISO Innovative Analytics, as well as various other efforts to introduce modern predictive modeling techniques, especially GLMs, into standard ISO ratemaking. Much of his career prior to ISO was spent as a research actuary involved in financial and statistical modeling at CIGNA P&C, with brief forays into regulation at the Massachusetts Insurance Department and reinsurance at Zurich Re North America, later Converium.