ACTUARIAL RESEARCH CLEARING HOUSE
1985 VOL. 2

Structured Credibility in Applications - Hierarchical, Multidimensional, and Multivariate Models

Gary Venter

Actuarial theory has traditionally accounted for the randomness inherent in a sample estimate by use of a credibility factor Z, where the final estimate U is a credibility weighted sum of the sample estimate X and some earlier estimate V, i.e., $U = Z\,X + (1 - Z)\,V$. For instance, in the case of class data, X could be the experience of an individual class and V the average of all classes combined.

The credibility factor Z has been calculated in various ways, and different choices for the previous estimate V have also been used. The least squares approach to credibility provides a mathematically well-grounded development of these elements in one specific setting.

Generally speaking, the previous estimate V is intended to represent the best estimate available without the particular observation X. The criteria for determining "best" usually involves informed judgment of some sort. Often V is taken to be the prior year's credibility weighted estimate. Using the overall group data as the previous estimate in a classification framework, is also prevalent. In some settings more than one previous estimate is used, each with a different credibility, e.g., $U = Z_0\,X + Z_1\,V + (1 - Z_0 - Z_1)\,W$, where V and W each have some claim on being a good estimate.

The use of informed judgment in selecting the prior estimate is one subjective aspect of the historical credibility analysis (there are others). The least

squares approach, on the other hand, can be carried out entirely from observations, without references to judgment, subjective probabilities, or prior beliefs. Because of this, non-Bayesian interpretations of this method are becoming understood, even though it was first developed in a Bayesian framework.

The least squares approach is developed for estimation within a classification framework, in which the pre-credibility estimate X_i is the observation for the ith class and the prior estimate V is the average of all classes. Thus, the credibility estimate U_i for the ith class is:

$$U_i = Z X_i + (1 - Z) V = Z X_i + (1 - Z) \sum_{j=1}^{n} X_j/n,$$

where Z can be expressed in the form P/(P + K). In the least squares approach this formula is generalized somewhat to allow for the possibility that different weights could apply to each class, that is, the credibility estimate is taken to be $U_i = \sum_{j=1}^{n} Z_j X_j$, or even more generally, $U_i = Y + \sum_{j=1}^{n} Z_j X_j$, i.e., a constant term is added. The problem is then formulated as finding the values for Y and the Z_j that give the optimal estimate U_i in the sense of least squares. In simple cases the answer turns out to be expressible in the original form $U_i = Z X_i + (1 - Z) \sum_{j=1}^{n} X_j/n$, i.e., all the Z_j turn out to be the same except for j = i, and the weights sum to unity.

The least squares approach can be put into a quite general framework. The problem is posed as estimating a random variable X_o as a linear combination of random variables $X_1, \ldots X_n$, i.e., $X_o \sim Y + \sum_{i=1}^{n} Z_i X_i$. Examples will be given below, but one application could be that the variables $X_1, \ldots, X_n$ each

represent the experience of a class for a single year, where several years are available for each class, and X_0 represents one of the classes for the next year. This is the original application of this theory. Developing it in this more general framework allows for extensions to be derived in a straightforward manner.

Three basic forms of extensions of the original application will be considered herein.

1. Credibility in a hierarchical or nested series of groups. An illustration is the situation where a class is a member of a group, and the group is incorporated in a cluster of groups. The class experience is credibility weighted against all the classes in the cluster but those classes in the same group receive greater weight. In other words, some credibility is given to the class experience, some to the local group, and the remainder to the average of all groups.

2. The simultaneous credibility estimation of several correlated variables. For instance, if class frequencies for different types of injury are correlated, then each would contain information that could be used to help estimate the other types. The class experience for each injury type would be credibility weighted against the other classes' experience for that injury type and the class experience for the other injury types.

3. A two-dimensional weighting. For example, the frequency for a class in a state would be weighted against that class's frequency in other states and the other classes' frequencies in that state.

The mathematical development of the P/(P + K) formulation for credibility for the original application follows. This is done in a fairly general framework, involving the correlations of the various random variables X_i with each other. This framework will provide the basis for the extensions of this approach to be developed.

To employ this framework, model assumptions made about the variables X_i must be used to express the expected values and covariances of the variables X_i in terms of postulated model parameters. From this covariance structure, expressions for the Z_i in terms of the model parameters can be derived.

Estimators for the model parameters from the data at hand can also be derived from the covariance relationships.

The covariances thus play a key role in the development of credibility mathematics. They link the model parameters to the data used to estimate these parameters, and to the credibility estimators derived from the parameters and the model assumptions.

Moreover, since different model assumptions may lead to the same covariance structure, just specifying the covariances may be a more general way to specify the model. On the other hand, it is sometimes difficult to get an intuitive

grasp of a model when it is so specified. Thus examples of model assumptions leading to different covariance structures will be given for each case discussed.

In the least squares set up, the values for Y and the Z_i in the above formula that give the optimal estimate are sought. Optional means in the sense of expected least squares, i.e., Y and Z_i are sought to minimize $E(X_o - Y - \sum_{i=1}^{n} Z_i X_i)^2$. It turns out that the Y and Z_i that do this can be expressed as:

$$Y = E(X_o) - \sum_{i=1}^{n} Z_i E(X_i)$$

where the Z_i are the solutions of the system of n equations:

$$\mathrm{cov}(X_o, X_j) = \sum_{i=1}^{n} Z_i \mathrm{cov}(X_i, X_j) \quad j = 1, \ldots, n.$$

This can be proved quite readily by taking the partial derivatives of the expectation to be minimized. In fact, the partial with respect to Y gives the first equation. Multiplying this by $E(X_j)$ and subtracting it from the partial with respect to X_j gives the second.

Thus, if the covariances are known, the Z_i can be found by solving this linear system by standard procedures. Then Y can be determined from the expected values. This is the first link in the credibility calculation: calculating the Z_i, and therefore estimating X_o, from the covariance relations. To make this useful, it will be necessary to have a way of calculating the covariances from the model assumptions. This will be addressed below.

Now it could be asked, if you know $E(X_o)$, why bother with this procedure at all? The answer to this becomes more specific when this approach is applied to a particular model; heuristically speaking, in most applications $E(X_o)$ denotes the global expectation; that is, it is the expected value of a class drawn at random from the collection of all classes, so $E(X_o) = E(X_1) = E(X_j) = E(X_n)$ is the overall average of all classes, and is thus the same for every class. The conditional expectation of X_o given all the observations $X_1,\ldots,X_n$ is more like what is being sought. As will be seen below, the credibility estimator approximates this conditional expectation, sometimes quite closely. In short, $E(X_o)$ is not the optimal estimator for X_o, i.e., it does not minimize the expected predictive error.

If the covariances and expectations needed to solve for Y and the Z_i are not known they can be estimated. However, the resulting Y and Z_i's are then no longer optimal to the extent that there is estimation error. Some adjustment can be made for this, although the theory behind common adjustments is incomplete. Since the quantities (i.e., variances) being estimated are of a global nature, there should be significantly less estimation error for them than for individual class estimates of the same quantities.

Some examples of the application of this theory should help clarify the above general discussion. First, some notation will be useful. Let $d(i,j) = 1$ if $i=j$ and $d(i,j) = 0$ otherwise. This indicator function is very useful when dealing with covariances. For instance, $cov(X_i, X_j) = d(i,j)s^2$ would say that the variance of each variable X_i is s^2 and any two different variables are uncorrelated.

A univariate one-dimensional single layer model

With this background, the first, and original, model-type considers class experience over time. $X_a(t)$ denotes the experience for class a at time t. Usually a ratio is being estimated, such as loss ratio, claim frequency, claim severity, or pure premium. Thus $X_a(t)$ is this ratio. The denominator is taken to be a known quantity (premium, exposure, claim count, etc.), and this will be denoted by $P_a(t)$.

To illustrate the first model-type, assume that the ratio for a particular class and time period can be decomposed as follows:

$$X_a(t) = m + R_a + T_a(t),$$

where R_a is the systematic (i.e., time independent) deviation of class a from the overall average m and $T_a(t)$ is the random deviation of $X_a(t)$ from its underlying mean. Further assume that $E(R_a) = 0$ and $E(T_a(t)) = 0$. That is, the systematic and random deviations are expected to average to zero over all classes and time periods. Different R's and T's are assumed to be independent random variables. To specify the variances of these variables, there are assumed to be two constants s and K such that $E(T_a(t)T_b(u)) = d(t,u)\,d(a,b)s^2/P_a(t)$, and $E(R_aR_b) = d(a,b)s^2/K$. The last two assumptions imply that $Var(T_a(t)) = s^2/P_a(t)$ and $Var(R_a) = s^2/K$.

Thus the random fluctuation term $T_a(t)$ has variance $s^2/P_a(t)$ which is inversely proportional to the denominator of $X_a(t)$. The quantity s^2 is just the constant of proportionality. The conditions where this inverse proportionality may be appropriate or inapproriate are discussed later. The

factor s is retained in the specification of var (R_a), but this is only for notational convenience. The variance s^2/K is usually calculated as a unit, then K backed out separately.

In this set up, the global expectation of $X_a(t)$ is m, even though there is a systematic deviation R_a, i.e., each class over time will display its own average.

Under these assumptions, what is the covariance of two observations $X_a(t)$ and $X_b(u)$? Since $cov(XY) = E(XY) - E(X)E(Y)$ and $E(X_a(t)) = E(X_b(u)) = m$, $cov(X_a(t)X_b(u)) = E(X_a(t)X_b(u)) - m^2$. Substituting the formula for the X's in terms of the R's and T's and multiplying out yields:

$$\begin{aligned}
&cov(X_a(t)X_b(u)) = m^2 + mE(R_a) + mE(R_b) + \\
&mE(T_a(t)) + mE(T_b(u)) + E(R_aR_b) + \\
&E(R_aT_b(u)) + E(R_bT_a(t)) + E(T_a(t)T_b(u)) - m^2 \\
&= E(R_aR_b) + E(T_a(t)T_b(u)) \text{ (by independence of R's and T's)} \\
&= d(a,b)s^2((1/K) + d(t,u)/P_a(t))
\end{aligned}$$

The credibility estimators that derive from this covariance structure are developed below. Other seemingly more general models will also be seen to produce this covariance structure, and so will produce the same credibility estimates.

Thus in summary, the first model-type assumes the following covariance structure:

$$\begin{aligned}
&E(X_a(t)) = m \\
&cov(X_a(t),X_b(u)) = s^2d(a,b)\ ((1/K) + d(t,u)/P_a(t))
\end{aligned}$$

This implies that the global expectation for each class and time period is the same, separate classes are uncorrelated, the covariance of two different time periods for a single class is s^2/K, and the variance of $X_a(t)$ is $s^2((1/K) + 1/P_a(t))$.

For this model specification, the optimal Y and $Z_a(t)$ are sought for the estimate of $X_g(0)$, the ratio for an unobserved time period for class g. The estimating relationship is:

$$X_g(0) = Y + \sum_{a,t} Z_a(t)\, X_a(t).$$

By the general least squares development above, the optimal Y and $Z_a(t)$'s are the solutions of the system:

$$Y = E(X_g(0)) - \sum_{a,t} Z_a(t)E(X_a(t))$$
$$\mathrm{cov}(X_g(0),X_h(u)) = \sum_{a,t} Z_a(t)\ \mathrm{cov}(X_a(t),X_h(u)),$$

where there is a copy of the last equation for every class h and observed time period u. By plugging in the covariance assumptions, this system can be converted to a system of equations for the unknown Z's in terms of the known P's and the two parameters s and K, and in fact s will drop out. Then the system can be solved for the Z's. This can be done by matrix methods, but for this model it will be possible to solve explicitly for the Z's, by straightforward algebra.

From the defining equations of the model-type (i.e., the covariance structure assumed) the system of equations becomes, after some algebra:

$$Y = m\ (1 - \sum_{a,t} Z_a(t))$$
$$s^2 d(g,h)/K = s^2\left(\sum_t Z_h(t)/K\right) + Z_h(u)s^2/P_h(u)$$

In this last equation many terms have dropped out because of the d function in the covariance assumption. Using a dot to denote the sum over an index, this equation then becomes:

$$P_h(u)d(g,h) = z_h(.)P_h(u) + Kz_h(u)$$

Since there is a copy of this for each h and u, sum over u to yield:

$$P_h(.)d(g,h) = z_h(.)P_h(.) + Kz_h(.), \text{ or}$$

$$z_h(.) = P_h(.)d(g,h)/(K + P_h(.))$$

Substituting this above yields:

$$z_h(u) = d(g,h)P_h(u)(1-(P_h(.)/(K + P_h(.)))/K$$

$$= d(g,h)P_h(u)/(P_h(.) + K)$$

This expresses the z's in terms of K and the known P's. Thus the optimal estimate becomes:

$$X_g(0) = Y + \sum_{h,u} z_h(u)X_h(u)$$

$$= m(1-\sum_{h,u} z_h(u)) + \sum_u (P_g(u)/(P_g(.) + K))\ X_g(u)$$

$$= m(1-P_g(.)/(P_g(.) + K)) + \sum_u P_g(u)X_g(u)/(P_g(.)+ K)$$

$$= m(K/(P_g(.) + K)) + \overline{X_g(.)}\ P_g(.)/(P_g(.) + K)$$

That is, $X_g(0)$ is estimated by a weighted average of the grand mean m and $\overline{X_g(.)}$, which is the weighted average ratio for class g, weighted in proportion to $P_g(t)$. The terms for the other classes have dropped out. Since g was any class, this same credibility formula applies to all classes, i.e., the K is the same, although the credibilities will be different.

The estimation of K will be discussed later. The above derivation illustrates how the credibility formula can be derived from the covariance structure. In the more general model-types discussed below, derivations are similar so less detail will be needed.

The above covariance structure was derived from a simple linear decomposition of $X_a(t)$. Another model, due to Buhlmann and Straub, that gets to this same covariance structure is based on conditioning. The model assumptions are given below and the above covariance structure is derived from these assumptions. This will demonstrate that the above credibility formula holds for these assumptions.

Each class a is assumed to carry an unobserved parameter V_a which determines the probability distribution of $X_a(t)$. The various V_i's are assumed to be independent and identically distributed random variables. Let $V = (V_1, V_2, \ldots)$. The following assumptions are made:

$$E(X_a(t)|V) = m(V_a)$$

$$Var(X_a(t)|V) = s^2(V_a)/P_a(t),$$

that is, there is a function $m(V_a)$ which will give the expected value for a class with parameter V_a, and this is independent of time, and there is a function $s^2(V_a)$ which will give the variance for a class with parameter V_a up

to the constant $P_a(t)$. Let $m = E(m(V_a))$. This does not depend on a because the V_i's are identically distributed. Let $s^2 = E(s^2(V_a))$, and $K = s^2/var(m(V_a))$. Then the covariance structure above holds as long as $X_a(t)$ and $X_b(u)$ are independent given V, for in that case:

$$cov(m(V_a), m(V_b)) = d(a,b)Var(m(V_a))$$

(independence of V_i)

$$\text{and } cov(X_a(t),X_b(u)|V)) = d(a,b)d(t,u)Var(X_a(t)|V)$$

(conditional independence of X's).

Thus:

$$cov(X_a(t), X_b(u)) = E\ cov(X_a(t),X_b(u)|V) + cov(E(X_a(t)|V),E(X_b(u)|V))$$

(by general principles of conditioning)

$$=E(d(a,b)d(t,u)Var(X_a(t)|V_a)) + d(a,b)\ Var(m(V_a))$$

$$=d(a,b)d(t,u)s^2/P_a(t)+d(a,b)s^2/K.$$

This model makes the notion of a global expectation that is the same for all classes with a conditional expectation that is unique to each class more precise for both the first and second moments. This model is also the paradigm of this type of analysis. In the more general models that follow, only the covariance structure will be specified, along with a simplified linear example of a model that will produce that covariance. Each could be represented in a more general conditioning framework, however.

A univariate one dimensional two layer model

The second covariance structure is the hierarchical model. In this, each class belongs to a unique group of classes. This can be generalized to groups of groups, etc. The random variables will be denoted as $X_{Aa}(t)$, and represent a ratio for class a in group A for time t. The assumed expections and covariances are:

$$E(X_{Aa}(t)) = m$$

$$\text{cov}\ (X_{Aa}(t),\ X_{Bb}(u)) = d(A,B)s^2((1/K)+d(a,b)((1/K_A)+d(t,u)/P_{Aa}(t)))$$

Here $P_{Aa}(t)$ is the denominator of $X_{Aa}(t)$. A model having this structure is $X_{Aa}(t) = m + Q_A + R_{Aa} + T_{Aa}(t)$, where the Q's, R's, and T's have global expectation zero and are mutually independent, and $E(Q_AQ_B) = d(A,B)s^2/K$, $E(R_{Aa}\ R_{Bb}) = d(A,B)d(a,b)s^2/K_A$, and $E(T_{Aa}(t)T_{Bb}(u)) = d(A,B)d(a,b)d(t,u)s^2/P_{Aa}(t)$.

In this model, Q_A represents the systematic departure of group A from the overall mean, R_{Aa} is the departure of class a from the group A mean, and $T_{Aa}(t)$ is the random fluctuation.

For a particular class, $X_{Gg}(0)$ is estimated as:

$$X_{Gg}(0)=Y+\sum_{A,a,t} Z_{Aa}(t)\ X_{Aa}(t).$$

The optimal weights under the general theory above are as follows:

$$Y = m(1 - \sum_{A,a,t} Z_{Aa}(t))$$

$$Z_{Aa}(t) = 0 \text{ except for } A=G.$$

For A = G:

Let $W_{Aa}(t) = P_{Aa}(t)/(P_{Aa}(.) + K_A)$

Then $Z_{Ga}(t) = d(a,g)\ W_{Gg}(t) + (1-W_{Gg}(.))W_{G.a}(t)/(W_G(.)+K/K_G)$

These formulas are derived in the same manner as for the single layer model. With some algebraic manipulation, the estimating equation can be written as: $X_{Gg}(0) = W_{Gg}(.)\ \overline{X_{Gg}(.)} + (1-W_{Gg}(.))m_G$, where $m_G = V_G M_G + (1-V_G)m$ with $V_G = W_G(.)/(W_G(.) + K/K_G)$ and $M_G = \sum_a W_{Ga}(.)\ \overline{X_{Ga}(.)}/W_G(.)$ Again the bar over the X's denotes the average weighted proportionally to the P's. This formulation shows that the hierarchical model can be considered to be a stepwise application of the simpler model. Several authors have discussed this model from a conditioning perspective, especially in the Scandinavian Actuarial Journal. For example, see Jewel [5] and Taylor [7]. The estimation of the K's is discussed later.

A multivariate one dimensional single layer model

Another covariance structure is provided by the estimation of several correlated variables, such as frequencies for different injury types. For this, let $X_{Aa}(t)$ denote the injury type A frequency for class a at time t. In this model each injury type will have its own grand mean, denoted by m_A. In fact, for the simple decomposition model, each term gets an additional subscript A as follows:

$$X_{Aa}(t) = m_A + R_{Aa} + T_{Aa}(t),$$

Again it is assumed that $E(R_{Aa}) = E(T_{Aa}(t)) = 0$, and the various R's and T's are mutually independent. To specify convariances, there are assumed to be constants s_A and r_{AB} such that $E(R_{Aa}R_{Bb}) = d(a,b)r_{AB}$, and $E(T_{Aa}(t),T_{Bb}(u)) = d(A,B)d(a,b)d(t,u)s_A^2/P_{Aa}(t)$. The first of these equations is the unique one

for this model. It allows the systematic departure terms for two different variables for a given class to be correlated.

The covariance structure is given by:

$$E(X_{Aa}(t)) = m_A$$

$$Cov(X_{Aa}(t), X_{Bb}(u)) = d(a,b)[r_{AB} + d(A,B)d(t,u)s_A^2/P_{Aa}(t)]$$

The estimating equation is then

$$X_{Gg}(0) = Y + \sum_{A,a,t} Z_{Aa}(t)X_{Aa}(t)$$

and the optimal Y and Z's turn out to be:

$$Y = m_G - \sum_{A,a,t} Z_{Aa}(t)\, m_A$$

$$Z_{Aa}(t) = 0 \text{ unless } a=g$$

$$Z_{Ag}(t) = P_{Aa}(t)(r_{AG} - \sum_{B} Z_{Bg}(.)\, r_{AB})/s_A^2$$

where the $Z_{Bg}(.)$ are the solutions of the system of equations:

$r_{GB} = \sum_{A} Z_{Ag}(.)r_{AB} + s_B^2 Z_{Bg}(.)/P_{Bg}(.)$ (one equation for each B).

Again the derivation of this result follows the general logic of the univariate model. Estimating the s_B and the r_{AB} will be discussed below. Note that in this model r_{AB}, which relates to the correlation between the expected frequencies for the injury types, must be estimated for each pair of injury types.

Another model with this same covariance structure can be defined following the Buhlmann-Straub approach, by establishing distinct functions m_A and s_A for each injury type. Thus:

$$E(X_{Aa}(t)|V) = m_A(V_a) \text{ and}$$
$$Var(X_{Aa}(t)|V) = s_A^2(V_a)/P_{Aa}(t).$$

Let $r_{AB} = Cov(m_A(V_a), m_B(V_a))$. Since the V_a are identically distributed, r_{AB} is not a function of a. Then the derivation of the covariance structure of the Buhlmann-Staub model outlined above can be used for this model to show that $cov(X_{Aa}(t), X_{Bb}(u)) = d(a,b)[r_{AB} + d(A,B)d(t,u)s_A^2/P_A(t)]$.

Univariate two dimensional single layer models

Finally, cross classification is considered. Here $X_{Aa}(t)$ may be a frequency for state A and class a at time t, for example. State and class are assumed to have separate effects that act independently. The expectation/covariance structure for this model-type is:

$$E(X_{Aa}(t)) = m$$
$$cov(X_{Aa}(t), X_{Bb}(u)) = s^2((d(A,B)/K)+(d(a,b)/J)+d(A,B)d(a,b)d(t,u)/P_{Aa}(t))$$

Specific models motivating this structure are discussed below.

The estimating equation is:

$$X_{Gg}(0) = Y + \sum_{A,a,t} Z_{Aa}(t)X_{Aa}(t)$$

Again, $Y = m(1 - \sum_{A,a,t} Z_{Aa}(t))$

Finding $Z_{Aa}(t)$ is somewhat more involved in this case. Let $W_{Aa}=P_{Aa}(.)/(P_{Aa}(.)+K)$. Then the following system of equations (one for each a) can be solved for the $Z_{.a}(.)$:

$$(P_{Ga}(.)/K)+(d(g,a)P_{.a}(.)/J)-(W_{Ga}P_{G.}(.)/K) - \sum_C W_{Ca}P_{Cg}(.)/J =$$
$$(1+P_{.a}(.)/J)Z_{.a}(.) - \sum_C W_{.c}Z_{.c}(.)/J$$

Then the $Z_{A.}(.)$ can be calculated by:

$$(1+P_{.a}(.)/J)Z_{A.}(.)=d(G,A)P_{A.}(.)/K+P_{Ag}(.)/J - \sum_C P_{Ac}(.)Z_{.c}(.)/J$$

Finally, $Z_{Aa}(t) = P_{Aa}(t)((d(A,G)/K+d(a,g)/J)-Z_{A.}(.)/K-Z_{.a}(.)/J)$

An example of this covariance structure is given by the following model:

$$X_{Aa}(t) = m + Q_A + R_a + T_{Aa}(t)$$

Here Q_A is the state departure component for state A and R_a is the class departure component for class a. It is assumed that the Q's, R's, and T's are mutually independent, each with an expected value over all classes, states, and time periods of zero, and that there are constants s, J, and K such that $E(Q_AQ_B) = d(A,B)s^2/K$, $E(R_aR_b) = d(a,b)s^2/J$, and $E(T_{Aa}(t)T_{Bb}(u)) = d(A,B)d(a,b)d(t,u)s^2/P_{Aa}(t)$. Essentially, s^2 is the variance factor for the random fluctuation term T, and given that, K expresses the variance across states and J across classes.

This model-type can be generalized to include interaction between class and territory. For example, the above model can be expanded to include an interaction term:

$$X_{Aa}(t) = m + Q_A + R_a + C_{Aa} + T_{Aa}(t).$$

C_{Aa} expresses the systematic deviation of the state-class cell from the sum of m plus the state component Q_A and the class component R_a. If it is assumed that

$E(C_{Aa}) = 0$ and

$E(C_{Aa}C_{Bb}) = d(A,B)d(a,b)s^2((1/L_A) + (1/M_a))$, then the resulting expectation/covariance structure is:

$$E(X_{Aa}(t)) = m$$

$$cov(X_{Aa}(t),X_{Bb}(u)) = s^2((d(A,B)/K) + (d(a,b)/J) + d(A,B)d(a,b)((1/L_A) + (1/M_a) + d(t,u)/P_{Aa}(t)))$$

This covariance structure is more general than the particular model illustrated. The resulting weights $Z_{Aa}(t)$ can be calculated as follows. Let

$W_{Aa} = P_{Aa}(.)/(1 + P_{Aa}(.)((1/L_A) + (1/M_a)))$

$U_{Aa} = W_{Aa}((1/L_G) + (1/M_g))$ (where G and g give the fixed class-territory combination being estimated)

$$D_A = -(W_{Ag}/J)-d(G,A)(U_{Ag} + W_{A.}/K) + \sum_a (W_{Aa}/(J+W.a))((W_{Ga}/K)+d(g,a)(U_{Ga}+W._a/J))$$

$$M_{BA} = -d(A,B)(1 + W_{A.}/K) + \sum_a W_{Ba}W_{Aa}/K(J + W.a)$$

Then the $Z_{A.}(.)$ are the solutions of the system of equations:

$$\sum_B M_{BA}Z_{B.}(.) = D_A \text{ (one equation for each A)}$$

From there, the $Z._a(.)$ can be calculated via:

$$Z._a(.)(J+W.a)/J = (W_{Ga}/K) + d(g,a)(U_{Ga}+W.a/J)-\sum_A W_{Aa}Z_{A.}(.)/K$$

Then $Z_{Aa}(.)$ follows by:

$$Z_{Aa}(.)/W_{Aa} = (d(G,A)/K) + (d(g,a)/J) + d(G,A)d(g,a)((1/L_G)+(1/M_a) - (Z_{A.}(.)/K) - Z._a(.)/J$$

and finally,

$$Z_{Aa}(t) = Z_{Aa}(.)P_{Aa}(t)/P_{Aa}(.).$$

The derivation of this result requires some fairly lengthy algebra, but it is a straightforward application of the methods discussed above for the one-dimensional model.

In contrast to the correlated variables model, the constants, J, K, L_a, M_A, etc. must be estimated for each state and class, but not for pairs of states or classes.

If this model were defined mutiplicatively, e.g., $X_{Aa}(t) = m(1 + Q_A)(1 + R_A) + C_{Aa} + T_{Aa}(t)$, with the other assumptions the same as above, the resulting covariances would be:

$$\text{cov}(X_{Aa}(t)X_{Bb}(u)) = s^2((d(A,B)m^2/K) + (d(a,b)m^2/J) + d(A,B)d(a,b)((1/L_A) + (1/M_a) + (s^2m^2/JK) + d(t,u)/P_{Aa}(t))).$$

The extra term m^2s^2/JK can be absorbed into either L_A or M_a by redefining the constants, and the original covariance structure is maintained, just with different parameter values. Thus, the same credibility formulas can be used for the multiplicative and additive models. That is, let $K^1 = K/m^2$ and $J^1 = J/m^2$ and either $M_a^1 = JKM_a/(JK + s^2m^2M_a)$ or $L_A^1 = JKL_A/(JK + s^2m^2L_A)$. Then the covariance equation for the multiplicative model is of the same form as that for the additive model, and the same credibility estimates apply to both.

In practice this model seems to make more sense with either the $1/L_A$ or the $1/M_a$ parameters assumed to be zero.

Multivariate and Multidimensional Model Comparison

Instead of using the above two-dimensional credibility model, the multivariate model could be tried in a two-dimensional situation. For example, the frequencies for the different states could each be interpreted as different variables, and the correlations measured. In the cross classified model with interaction, where $1/M_a = 0$, for example, the decomposition formulas for the two models will look the same if $m_A = m + Q_a$ and $R_{Aa} = R_a + C_{Aa}$ are taken.

The covariances will not be the same, however. In the multivariate model m_A is a constant, where $m + Q_A$ is a random variable. This provides an extra element of variance, the s^2/K term, in the two-dimensional model.

Also, by taking $r_{AB} = s^2/J + s^2 d(A,B)/L_A$, the $E(R_{Aa} R_{Bb})$ terms of the two models will be equated. However, doing this would require estimating each r_{AB} separately, where in the two-dimensional model only s, J, and the L_A need be estimated. An important part of the structure is being ignored, so information is lost.

Estimation of Parameters

The final link in the credibility chain is to use the model assumptions, namely the covariance structure, to develop formulas to estimate the model parameters.

Natural unbiased estimators of some of the parameters needed in the various models above can be estimated from weighted sums of squared differences of the observed data. The remaining parameters can be developed as combinations of those so estimated.

The procedure in general will be to write down the sums of squared differences most suggested from the model, then use the expectation/covariance structure of the model to evaluate the expected values of these estimators. From these expected values, the desired unbiased estimators can be calculated.

While this is a natural and straightforward way to estimate variances, it is probably not optimal. It may be appropriate to temper with the resulting estimates judgmentally in many cases.

To illustrate the method, some details of the procedure will be shown for the simplest model, but only more summarized outlines will be given for the more general models.

For the univariate one-dimensional single layer model, then, it is assumed that observations $X_a(t)$ are available for $a = 1,\ldots,N$ and $t = 1,,,,n$. Let:

$$\overline{X_a} = \sum_t P_a(t)\ X_a(t)/P_a(.)$$

$$\bar{X} = \sum_{a,t} P_a(t)X_a(t)/P_\cdot(\cdot)$$
$$D_1 = \sum_{a,t} P_a(t)(X_a(t) - \bar{X}_a)^2$$
$$D_2 = \sum_{a,t} P_a(t)(X_a(t) - \bar{X})^2.$$

To evaluate $E(D_1)$ and $E(D_2)$, note that from the covariance structure, $E(X_a(t)X_b(u)) = m^2 + s^2 d(a,b)((1/k)+d(t,u)/P_a(t))$.

Also recall that in general,

$$(\sum_i Y_i)^2 = \sum_{i,j} Y_i Y_j$$

The first step is to evaluate $E(\bar{X}_a^2)$, $E(\bar{X}^2)$, $E(X_a(t)\bar{X}_a)$, and $E(\bar{X}_a\bar{X})$. Thus,

$$P_a(\cdot)^2 E(\bar{X}_a^2) = \sum_{t,u} E(P_a(t)P_a(u)X_a(t)X_a(u))$$
$$= \sum_{t,u} P_a(t)P_a(u)(m^2 + s^2((1/K) + d(t,u)/P_a(t)))$$
$$= (m^2 + s^2/K)\sum_{t,u} P_a(t)\, P_a(u) + \sum_t s^2 P_a(t)^2/P_a(t)$$
$$= (m^2 + s^2/K)\sum_t P_a(t)\, P_a(\cdot) + \sum_t s^2 P_a(t)$$
$$= (m^2 + s^2/K)\, P_a(\cdot)^2 + s^2 P_a(\cdot)$$

Thus $E(\bar{X}_a^2) = m^2 + s^2/K + s^2/P_a(\cdot)$

Similar algebra will yield:

$$E(\bar{X}^2) = m^2 + (s^2/K)_a(P_a(\cdot)/P_\cdot(\cdot))^2 + s^2/P_\cdot(\cdot)$$
$$E(\bar{X}X_a(t)) = m^2 + (s^2/K)(P_a(\cdot)/P_\cdot(\cdot) + s^2/P_\cdot(\cdot)$$
$$E(X_a(t)\bar{X}_a) = m^2 + (s^2/K) + s^2/P_a(\cdot)$$

Then:

$$
\begin{aligned}
E(D_1) &= \sum_{a,t} P_a(t)E(X_a(t)^2 - 2X_a(t)\bar{X}_a + \bar{X}_a^2) \\
&= \sum_{a,t} P_a(t)(m^2 + s^2/K + s^2/P_a(t) \\
&\qquad -2m^2 - 2s^2/K - 2s^2/P_a(.) \\
&\qquad +m^2 + s^2/K + s^2/P_a(.)) \\
&= s^2\sum_{a,t} P_a(t)(1/P_a(t) + 1/P_a(.) - 2/P_a(.)) \\
&= s^2(Nn + N - 2N) = N(n-1)s^2
\end{aligned}
$$

Thus $D_1/N(n-1)$ is an unbiased estimator of s^2.

Also:

$$
\begin{aligned}
E(D_2) &= \sum_{a,t} P_a(t)E(X_a(t)^2 - 2X_a(t)\bar{X} + \bar{X}^2) \\
&= \sum_{a,t} (P_a(t)(m^2 + s^2/K + s^2/P_a(t) \\
&\qquad -2m^2 - 2(s^2/K)(P_a(.)/P.(.) - 2s^2/P.(.)) \\
&\qquad +m^2 + (s^2/K)\sum_a (P_a(.)/P.(.))^2 + s^2/P.(.)) \\
&= \sum_{a,t} P_a(t)(s^2/K)(1-2P_a(.)/P.(.) + \sum_a (P_a(.)/P.(.))^2) \\
&\quad + \sum_{a,t} P_a(t)s^2(1/P_a(t) - 2/P.(.) + 1/P.(.)) \\
&= (s^2/K)\ (P.(.) - 2\sum_a P_a(.)^2/P.(.) + \sum_a P_a(.)^2/P.(.)) \\
&\quad + s^2(Nn - 2 + 1) \\
&= (s^2/K)(P.(.) - \sum_a P_a(.)^2/P.(.)) + s^2(Nn - 1)
\end{aligned}
$$

Thus $(D_2 - (Nn-1)D_1/N(n-1))/(P.(.) - \sum_a P_a(.)^2/P.(.))$ is an unbiased estimator of s^2/K. From this, K can be estimated. While the result is a natural estimator it is not in general unbiased.

While a fair amount of algebra is involved, the above calculation proceeds in a straightforward manner from the covariance structure. The same will be the

case for the other models. Estimation of the global expectation m will be discussed separately.

This development assumes that every class has exposure in every time period, but the estimation can proceed without this assumption. This is illustrated below, for the hierarchical model.

Hierarchical Model

It will be convenient to adopt the notation I(P) = 0 if P = 0 and I(P) = 1 otherwise. This will be used to count the number of cells with non-zero exposure. For example, $\sum_t I(P_{Aa}(t))$ would give the number of years of experience for class a in state A. M will be used to denote the number of A's, N the number of a's, and n the number of t's. For the two layer model, Heckman developed the following estimators:

$$D_1 = \sum_{A,a,t} P_{Aa(t)}(X_{Aa}(t) - \overline{X_{Aa}})^2$$
$$D_{2A} = \sum_{a,t} P_{Aa}(t)(X_{Aa}(t) - \overline{X_A})^2$$
$$D_3 = \sum_{A,a,t} P_{Aa}(t)(X_{Aa}(t) - \overline{X})^2,$$

which have the expected values:

$$E(D_1) = s^2\left(\sum_{A,a,t} I(P_{Aa}(t)) - \sum_{Aa} I(P_{Aa}(.))\right)$$
$$E(D_{2A}) = S^2(P_{A.}(.)\ (1 - \sum_a (P_{Aa}(.)/P_{A.}(.))^2)/K_A + \sum_{a,t} I(P_{Aa}(t)) - I(P_{A.}(.)$$
$$E(D_3) = s^2\left(\sum_{A,a,t} I(P_{Aa}(t)) - 1 + (P..(.)/K)(1 - \sum_A (P_A(.)/P..(.))^2) + \right.$$
$$\left. \sum_A (P_{A.}(.) - \sum_a (P_{Aa}(.)/P_{A.}(.))^2)/K_A\right)$$

Thus, D_1 provides an unbiased estimate of s^2, D_{2A} of s^2/K_A, and D_3 can be used to estimate K from K_A and s^2.

The expected values shown can be derived from the covariance structure in the same manner as for the basic model.

Correlated Variables Model

For the multivariate model, let:

$$D_A = \sum_{a,t} P_{Aa}(t)(X_{Aa}(t) - \overline{X}_{Aa})^2$$

$$D_{AB} = \sum_{a,t} (P_{Aa}(t)P_{Ba}(t))(X_{Aa}(t) - \overline{X}_A)(X_{Ba}(t) - \overline{X}_B)$$

By direct calculation from the covariance assumptions,

$$E(D_A) = s^2_A(n-1)N$$

$$E(D_{AB}) = r_{AB}(P_{A.}(.) + P_{B.}(.) - \sum_a P_{Aa}(.)^2/P_{A.}(.) - \sum_a P_{Ba}(.)^2/P_{B.}(.)) + 2s^2_A d(A,B)(Nn-1).$$

These provide unbiased estimators of both r_{AB} and s^2_A.

Two-Dimensional Models

For the two dimensional models, let:

$$D_1 = \sum_{A,a,t} P_{Aa}(t)(X_{Aa}(t) - \overline{X}_{Aa})^2$$

$$D_{2A} = \sum_{a,t} P_{Aa}(t)(X_{Aa}(t) - \overline{X}_{A.})^2$$

$$D_2 = \sum_A D_{2A}$$

$$D_{3a} = \sum_{A,t} P_{Aa}(t)(X_{Aa}(t) - \overline{X}_{.a})^2$$

$$D_3 = \sum_a D_{3a}$$

$$D_4 = \sum_{A,a,t} P_{Aa}(t)\ (X_{Aa}(t) - \bar{X})^2$$

Then the techniques above will yield:

$$E(D_1) = s^2MN(n-1)$$
$$E(D_2) = s^2(M(Nn-1) + (P..(.) - \sum_{Aa} P_{Aa}(.)^2/P_{A.}(.))/J)$$
$$E(D_3) = s^2(N(Mn-1) + (P..(.) - \sum_{Aa} P_{Aa}(.)^2/P.a(.))/K)$$

These can be used to estimate s^2, J, and K.

In the case of the model with interaction terms,

$$E(D_1) = s^2MN(n-1)$$
$$E(D_{2A}) = s^2(U_A/J + U_A/L_A + \sum_a R_{Aa}/M_a + Nn-1)$$
$$E(D_{3a}) = s^2(V_a/K + V_a/M_a + \sum_A Q_{Aa}/L_A + Mn-1)$$
$$E(D_4) = s^2(W_1/K + W_2/J + \sum_A C_A/L_A + \sum_a F_a/M_a + MNn-1)$$

where $U_A = P_{A.}(.) - \sum_a P_{Aa}(.)^2/P_{A.}(.)$

$V_a = P_{.a}(.) - \sum_A P_{Aa}(.)^2/P.a(.)$

$R_{Aa} = P_{Aa}(.) - P_{Aa}(.)^2/P_{A.}(.)$

$Q_{Aa} = P_{Aa}(.) - P_{Aa}(.)^2/P_{.a}(.)$

$W_1 = P..(.) - \sum_A P_{A.}(.)^2/P..(.)$

$W_2 = P..(.) - \sum_a P_{.a}(.)^2/P..(.)$

$C_A = P_{A.}(.) - \sum_a P_Aa(.)^2/P..(.)$

$F_a = P_{.a}(.) - \sum_A P_{Aa}(.)^2/P..(.)$

These equations are enough to estimate all but one parameter in a simultaneous system.

It may generally be more practical to set either the $1/M_a$ or the $1/L_A$ parameters to zero. Then D_1, D_4, and either D_2 and the D_{3a} or D_3 and the D_{2A} will be enough to estimate the remaining parameters.

Estimation of m

The overall mean would seem to be the easiest parameter to estimate. However, since it is so easy, more can be done in the way of deriving an optimal estimate. Because of this, the formulas for estimating the means may seem as complex as those for the other parameters.

In complex structures with auxiliary data at hand, the overall mean might be estimated outside of the model used to estimate specific classes, and this may be entirely legitimate. However, if it is to be estimated as a linear function of the observations X_i, the following approach is indicated.

The general credibility formula $X_o = Y + \sum_i Z_i X_i$ has been seen to require $Y = E(X_o) - \sum_i Z_i E(X_i)$, and thus can be written as $X_o = E(X_o) + \sum_i Z_i (X_i - E(X_i))$. From this it can be seen that the credibility estimate is unbiased. If the expected values themselves are to be estimated as weighted averages of the X_i, then this formula can be expressed as:

$$X_o = \sum_i Z_i X_i \text{ (different Z's)}$$

i.e., no constant term is needed. Thus, in the case that the expected values of the X_j are estimated as weighted averages of the observations, the credibility estimator is an unbiased linear estimate with no constant term.

If the Z_i's could be found that produced the optimal unbiased no constant estimate, in the sense of least squares, then equating this estimator with the

original credibility estimator would yield an implicit expression for the global mean needed in the original estimator. This would then be the best linear estimate of the global mean, in the sense that it produces the best least square credibility estimate of X_o.

To find the optimal no constant unbiased linear estimator for X_o, the technique of Lagrangian multipliers is used. This technique provides maximums or minimums of functions subject to a side condition, or constraint. In this case the function to be minimized is $E(X_o - \sum_i Z_i X_i)^2$ and the constraint is unbiasedness, i.e., $E(X_o) - \sum_i Z_i E(X_i) = 0$. The Lagrangian multiplier technique is to find the Y and Z_i that minimize the following function:

$E(X_o - \sum_i Z_i X_i)^2 - 2Y(E(X_o) - \sum_i Z_i E(X_i))$. Then the Z_i so determined will minimize the original function subject to the side constraint.

Taking partial derivatives of this function with respect to Y and the Z_i yields the following system of equations:

$$E(X_o) = \sum_i Z_i E(X_i)$$
$$\mathrm{cov}(X_o, X_j) = \sum_i Z_i \mathrm{cov}(X_i X_j) + Y E(X_j)$$

This system can be solved for the Z's in terms of the model parameters for each of the model-types discussed earlier, just by plugging in the covariances and algebraic manipulation.

Paralleling the original development for the basic model, the first equation implies $Z_.(.) = 1$. The second yields:

$$Z_h(.) = P_h(.)(d(g,h) - KYm)/(K + P_h(.))$$

Summing this over h will provide:

$$Ym = 1/(P_g(.) + K) \sum_h (P_h(.)/(P_h(.) + K))$$

After some algebra the estimate of $X_g(0)$ is:

$$X_g(0) = (K/(P_g(.) + K))(\sum_h \overline{X_h(.)} P_h(.)/(P_h(.) + K)) \div \sum_h P_h(.)/(P_h(.) + K) + \overline{X_g(.)}\ P_g(.)/(P_g(.) + K)$$

Equating this with the original estimate for this model yields:

$$m = (\sum_h \overline{X_h(.)} P_h(.)/(P_h(.) + K)) \div \sum_h P_h(.)/(P_h(.) + K)$$

This can be seen to be a weighted average of the observations, where the weight is proportional to the credibility $P_h(.)/(P_h(.) + K)$.

Referring now to the hierarchical model, the mean to be used for a group of classes, before looking outside the group, was seen to be just such a weighted average, i.e., it is the weighted average of the class means, where the weights are proportional to the class credibility. The mean of all the groups was not specified before, but it can be estimated by the above methodology. The result is a weighted average of the group means, where group A gets weight proportional to $W_A(.)/(W_A(.) + K/K_A)$. In fact, this weight is the credibility the group mean received in the overall weighting process. So again the overall mean is estimated as a weighted average of the cases, with the weights proportional to the credibility.

For the multivariate and multidimensional models the same sort of analysis can be carried out, but no closed form formulas for the means have been

developed. Formulas can be given in matrix notation. However, if this approach to estimating the mean is to be taken, it may be easier to develop the credibility formulas directly from the no constant approach without considering the mean explicitly.

Relation to Bayesian Analysis

The credibility estimates above were developed by minimizing the expected value of the square of the difference between the linear estimate $Y + \Sigma Z_i X_i$ and the variable X_0 being estimated. The Bayesian predictive mean, $E(X_0 / X_1 \ldots X_n)$, has been shown to be the unique function of any form, i.e., not just linear, of $X_1 \ldots X_n$ that minimizes the expected squared estimation error. A proof of this can be found in DeVlyder.

Thus, the credibility estimation error can be considered to have (at least) two components: the fundamental randomness component that even the optimal Bayesian estimate possesses, and a "linearization error," which is the additional error, if any, that arises from restriction to the class of linear estimators.

When the Bayesian estimate can be expressed as a linear function of the observations X_i, there is of course no linearization error, and the credibility estimate is equal to the Bayesian estimate. Jewell has shown that this will be the case for a fairly wide range of distributions, but for other distributions the linearization error can be significant. Many of the heavily skewed distributions that arise in Property and Casualty insurance applications fall into the latter category.

For both credibility and Bayesian estimates an additional source of error arises from the estimation of parameters. In fact, the Bayes estimate is optimal only if the joint distribution of the X_j's is known, and the optimum linear property of the credibility estimate holds only if the required moments are known. The question of whether Bayes estimation or least squares credibility will be more accurate may hinge on whether the additional approximations needed to estimate the joint distribution function of the X_j's instead of just the two moments needed for credibility will produce more error than the linearization error inherent in credibility estimates. This is currently an open issue in many practical settings.

Stein estimators provide another alternative. Estimation there is viewed as strictly a function of the data, i.e., the optimal estimate knowing just the data is sought. In some cases this approach has been shown to improve credibility estimates. By interpreting the Stein estimate as a credibility formula, an adjustment to the credibility factor can be derived that compensates for error in the estimation of K.

Stratifying the population is an alternative in some situations. Each of the strata would be more likely to approximate the requirements for no linearization error. However, more parameters would have to be estimated. This is the approach currently under investigation in Workers Compensation insurance.

Another alternative, to be discussed further below, is to use credibility on transformed values of the variables, then apply the inverse transformation to the estimate. This may reduce linearization error, but could magnify the error in estimation of parameters.

To compute Bayes estimates in the case of the simplest credibility model is more involved than it may seem, in that two distinct prior distributions may be needed, one for the mean and one for the variance. This is illustrated in the example below.

Consider the negative binomial distribution with parameters (y,q) in the form $Pr(N=n)=q^{y}(1-q)^{n}y(y+1)...(y+n-1)/n!$, where y and q are positive and q is below 1. Then $E(N)=y(1-q)/q$ and $Var(N)=y(1-q)/q^{2}$.

Suppose for class a for time t the number of exposure units is $P_a(t)$, the frequency is $X_a(t)$ and the number of claims $N_a(t)=X_a(t)P_a(t)$ is negative binomial distributed with parameters $(y_a P_a(t),q_a)$. Let V_a denote the pair (y_a,q_a). Then $E(X_a(t)/V_a)=E(N_a(t)/V_a)/P_a(t)=y_a(1-q_a)/q_a=m(V_a)$ and $Var(X_a(t)/V_a)= Var(N_a(t)/V_a)/P_a(t)^2=y_a(1-q_a)/q_a^2 P_a(t)=s^2(V_a)/P_a(t)$. Under these definitions of the functions m and s the conditions for the Buhlmann-Straub formulation are satisfied.

To use Bayes estimation instead of credibility, the joint density functions for y_a and q_a would be needed. This could be simplified to a single variable problem by assuming $q_a=q$ is a constant for all classes, or by taking q_a to be a function of y_a. Such assumptions are common in risk theory applications generally (e.g., see Patrik and John for an example of the former and Meyers and Heckman for the latter). However, the credibility model does allow for the more general case. The limited data currently available for Workers Compensation gives little support for a functional relationship between q's and y's. In fact scatter diagrams of class sample means and variances tend to support independence of q's and y's.

Relation Between Variance and Exposure

In the negative binomial example above, the variance of $X_a(t)$ was found to be inversely proportional to $P_a(t)$, with constant $y_a(1-q_a)/q_a$, consistent with the assumptions of the basic credibility model. The applicability of the inverse variance assumption is explored further in this section.

Let $X_a(t)=N_a(t)/P_a(t)$, where $P_a(t)$ is a constant, the "exposure," and $N_a(t)$ is a random variable, such as total number or dollar amount of claims. Denote $E(X_a(t))$ as m_a, so $E(N_a(t))=m_aP_a(t)$. While not denoted as such, these expectations could be understood as conditional on some parameter. The notation assumes that the expected value of $X_a(t)$ does not depend on t. Note that $\text{Var}(X_a(t))= \text{Var}(N_a(t))/P_a(t)^2$.

Consider the case where there is a proportional relation between the mean and variance of N, i.e., $\text{Var}(N_a(t))= cE(N_a(t))$. Then $\text{Var}(X_a(t))=cE(N_a(t))/P_a(t)^2 = cm_a/P_a(t)$, i.e., the inverse relation holds.

It turns out, as discussed below, that the proportional relationship for the mean and variance of N is a reasonable assumption for both the number and aggregate amount of claims. Thus, taking appropriate measures for P, this would lead to the inverse relation for variance for frequency, severity, pure premium, and loss ratio.

What makes the proportional relationship between variance and mean a reasonable assumption is that the proportion does not depend on volume. Thus, let N be either the number or aggregate amount of claims and let K = Var(N)/E(N). Let M be a similar random variable, independent of N, such that Var(M)/E(M) is also equal to K. Then Var(N + M) = Var(N) + Var(M) = KE(M) + KE(N) = KE(N+M). Thus the large volume of exposure maintains the same relation of variance to mean as the parts.

Thus the inverse relationship between variance and exposure is a reasonable assumption in many situations. Now a situation where it does not hold is explored.

Again, let the parameter V_a for class a consist of the pair y_a, q_a. The number of claims for time period t is determined as follows:

A random variable H is drawn from a specific distribution with $E(H) = y_a q_a$ and $Var(H) = y_a q_a^2$, perhaps a gamma distribution. Then $N_a(t)$ is drawn from a Poisson distribution with mean $HP_a(t)$. Thus $E(N_a(t)/V_a) = EE(N_a(t)/H) = P_a(t) y_a q_a$ and $Var(N_a(t)/V_a) = E(Var(N_a(t)/H)) + Var(E(N_a(t)/H)) = E(HP_a(t)) + Var(HP_a(t)) = P_a(t) y_a q_a + P_a(t)^2 y_a q_a^2$. (If H is in fact gamma distributed, $N_a(t)$ turns out to follow a negative binomial, this time with parameters $(y_a, 1/(1+P_a(t)q_a))$. Note that the dependence on P is now in the other parameter than that from the earlier negative binomial example.)

The moment formulas for N yield:

$E(X_a(t)/V_a) = y_a q_a$

$Var(X_a(t)/V_a) = y_a q_a^2 + y_a q_a / P_a(t)$. The variance of N is not proportional to the mean, and thus the variance of X is not inversely proportional to P. This occurs because the element of variance that arises from the selection of H is not related to P.

For this model it is still possible to define $m(V_a) = y_a q_a$, but now the conditional variance is a function $S_1(V_a) + S_2(V_a)/P_a(t)$. This formulation could be used in developing credibility formulas. The estimation of $E(S_1(V_a))$ and $E(S_2(V_a))$ may require regression.

There may be difficulties in determining whether this model is more appropriate than the earlier negative binomial model. The functions $s^2(V_a)$, $S_1(V_a)$, and $S_2(V_a)$ may take on very different values for different classes, so comparing sample variances among classes is not definitive.

In other words, if the sample frequency variances of classes with large values of P were found to be not too much smaller than the frequency variances of classes with small values of P, one might think that $S_1(V_a)$ tended to be fairly large. However, such a finding is not inconsistent with $S_1 = 0$. The classes with large P could also be the classes with large values of $s^2(V_a)$, for example. This would violate the assumption of iid V_a's, however, and thus would also change the credibility calculation.

An area in which the implications of the two models differ is in the experience over time of single classes. Both models predict that for a given class, the years with higher P will tend to stay closer to the class average, but the dependence of this tendency on P will be greater for the first ($S_1=0$) model. This difference may be subtle, however, and it is not clear how much data is needed to identify it.

The former effect has been seen in Workers Compensation data, i.e., the larger classes do not have proportionally smaller frequency variances than do the smaller ones. Groups of large classes tend to indicate higher K's, and hence tighter credibility requirements, than do groups of smaller classes. It could be that the larger classes are inherently more variable and $S_1=0$, but the alternate hypothesis also seems plausible.

If S_1 is positive, then S_2 will probably be estimated as a lower number than the s^2 of the other model, so the credibility of the large classes may rise. This would be consistent with an effect that has been noticed: the predictive accuracy of the credibility models seems to improve by instituting a finite maximum for K. This ends up giving the larger classes higher credibility than the model assumptions would imply, and its efficacy suggests a failure of some hypothesis, perhaps the inverse relation between exposure and payroll.

Another suspect hypothesis is time homogenity, i.e., that the class averages are consistent over time. For Workers Compensation class frequency, predictive accuracy was improved by using, for each class, only the number of years of data needed to produce 95% credibility for that class, based on the K determined from the model. This is an ad hoc adjustment for time heterogenity, but it seems useful.

Credibility Formulas Without Inverse Proportionality

As above, consider the Buhlmann-Straub formulation, but without the inverse variance assumption. For convenience, specify the covariance relationship as follows:

$$\text{Cov}\ (X_a(t), X_b(u)) = s^2 d(a,b)(1 + d(u,t)(J + K/P_a(t)))$$

The mixed Poisson case above is an example of this form. Carrying out the usual development with this relation yields the estimate:

$$X_g(0) = m(1 - z_g(.)) + \sum_t z_g(t) X_g(t),$$

where $Z_g(u) = W_g(u)/(1 + W_g(.))$ and $W_g(u) = P_g(u)/(JP_g(u) + K)$.

For the models discussed earlier, a credibility factor could be applied to the class average experience over the observed time periods. In this formulation a separate factor is needed for each period. In the special case where all the years have equal exposure for a class, the formula simplifies somewhat. The estimate becomes:

$$X_g(o) \approx m(1-z_g(.)) + z_g(.)\overline{X_g(.)},$$

with $Z_g(.) = P_g(.)/(P_g(.)(1 + J/n) + K)$. Here n is the number of years observed.

To estimate s^2, J and K, let:

$$D_{1a} = \sum_t P_a(t)(X_a(t) - \overline{X_a(.)})^2$$

$$D_2 = \sum_{a,t} P_a(t)(X_a(t) - \overline{X(.)})^2$$

Then the covariance assumption yields, after some algebra,

$$E(D_{1a}) = (n-1)Ks^2 + Js^2(\sum_t P_a(t)^2 - P_a(.)^2)/P_a(.)$$

From this, the quantities $(n-1)Ks^2$ and Js^2 can be estimated as the intercept and slope of a regression. Then s^2 can be estimated by use of:

$$E(D_2) = Ks^2(Nn-1) + Js^2(\sum_{a,t} P_a(t)^2 - P.(.)^2)/P.(.) +$$
$$s^2 \sum_a (P_a(.)^2 - P.(.)2)/P.(.).$$

This will then yield estimates of J and K.

Non-Linear Exponential Families

If X is distributed according to a frequency in a linear exponential family, the credibility estimate has been shown to equal the Bayes estimate, under standard normalcy assumptions. This is essentially because for these families the observed mean is a complete sufficient statistic, and so an unbiased linear function of this mean will be the best unbiased estimator.

If instead of X some transformation of X, T(X), follows a linear exponential family, then the mean of the transformed observations is a complete sufficient statistic. For several transformed linear exponential families, the inverse transformation of the transformed mean will be proportional to the (untransformed) mean. In these cases a constant times the inverse transformation of the credibility estimate of the transformed variable will then be an unbiased estimate, and hence the best unbiased estimate, of X. The constant is not necessarily determined by these considerations, but could be specified by requiring overall balance to a particular level.

Some examples will be useful.

Suppose M is normal in $N;S^2$ and X/M is lognormal in $M;T^2$. Then the predictive distribution of X given observations $X_1, \ldots, X_n$ can be shown to be lognormal in: $(T^2N + S^2\Sigma \ln x_i)/(T^2 + S^2n)$; $T^2+T^2S^2/(T^2+nS^2)$, with expected value of: $\exp(T^2(n+K+1)/2(n+K))\ \exp\ (N(1-Z)+Z\overline{\ln X_i})$, where $K = T^2/S^2$ and $Z = n/(n+K)$. Thus the posterior expected value is proportional to the invers

transformation of the credibility estimate of the transformed variable. The standard credibility estimate, for comparison, is given by $Z_1\overline{X}_i + (1-Z_1)\exp(N+(S^2+T^2)/2)$, where $Z_1 = n/(n+K_1)$ and $K_1 = \exp(S^2)(\exp(T^2)-1)/(\exp(S^2)-1)$.

For another example, suppose B is distributed inverse transformed gamma in s;a;c with density $f(b)=(a/c)\exp(-(c/b)^a)(c/b)^{sa+1}/\Gamma(s)$, and moments $E(B^n) = c^n\Gamma(s-(n/a))/\Gamma(s)$, and X/B is distributed transformed gamma in r;a;B with density $g(x)=(a/B)(x/B)^{ar-1}\exp(-(x/B)^a)/\Gamma(r)$, with moments $E(X^n/B)=B^n\Gamma(r+(n/a))/\Gamma(r)$. These moment formulas hold for all real numbers n as long as n exceeds -ar for the transformed gamma or n is below as for the inverse transformed gamma.

The Bayes estimate can then be shown to be $J((1-Z)E(X^a)+Z\overline{X_i^a})^{1/a}$, where $E(X^a) = rc^a/(s-1)$, $Z = n/n+K$, $K=(s-1)/r$, and $J=(n+K)^{1/a}\Gamma(s+nr-1/a)\Gamma(r+1/a)/\Gamma(s+nr)\Gamma(r)$. Again, this is proportional to the inverse transformation of the credibility estimate of the transformed variable.

This approach to credibility requires that the transformation be known, i.e., the log or some known power of the variable should be felt to follow a distribution from a linear exponential family. The credibility estimate can then be done on the transformed data and the inverse transformation applied. The constant of proportionality can then be arrived at by finding the factor needed to balance to the observed mean of all classes combined.

Interpretation of Models

For those who have philosophical difficulty with the conditioning models, the decomposition, or variance component models used above may provide a useful alternative viewpoint for the interpretation of credibility formulas.

Having difficulty with the conditioning formulation is not limited to those with a frequency view of probability. The line of reasoning below is sometimes heard from advocates of a subjective uncertainty perspective.

The subjective viewpoint is sometimes categorized as asserting that probability is orderly opinion. This may be somewhat strong, especially if it is taken to imply that all other forms of opinion are disorderly.

For purposes of discussion, subjective probability will be defined as a three-way relationship among a statement, an observer, and a number from 0 to 1 such that the number represents the observer's degree of confidence in the statement.

For this definition, a statement is assumed to be, or at least seems to be, meaningful, but it does not necessarily have to be either true or false. Thus the statement, "Hamlet was a Gemini" can be assigned subjective probabilities by various observers.

While some probabilities may be different from observer to observer, others may be quite stable. In other words, "subjective" does not necessarily imply "unreliable." As an example, "the trillionth digit of e is a 7" would be assigned a probability of .1 by many observers.

Taking the observer out and defining objective probabilities becomes problematic for some analysts. There is a point of view that would assign an objective probability of 1 to all true statements, 0 to all false, and not allow objective probability to apply at all to other statements. Thus, "Hamlet was a Gemini" would not be eligible for objective probability, while

"The trillionth digit of e is a 7" has objective probability of either 0 or 1.

Under this viewpoint, statements about the future have no special status. They are still either true, false, or neither. Thus "It will rain on my birthday" objectively would have probability either 0 or 1, depending on whether the statement is true or false. This does not rely on the truth or falsity being determinable in advance, even in principle.

From this point of view, it becomes difficult to interpret statements like: "A has a Poisson distribution for claim frequency, with the parameter following a gamma distribution." The problem is not with the gamma, because that can be interpreted as someone's uncertainty about something (the parameter). The problem is with the Poisson, which sounds like an objective probability, and thus should be either 0 or 1.

This philosophical difficulty may not be a problem for credibility applications, because the formulas themselves have simpler, less problematic interpretations. Resolving this difficulty would nonetheless help clarify thinking and discussions about the concept of probability and its role in credibility models.

References

(1) Buhlmann, H.. and Straub, E., "Credibility for Loss Ratios," English transaction by C.E. Brooks, Zurich.

(2) DeVlyder, F. "Introduction to the Actuarial Theory of Credibility," English translation by Charles A. Hachemeister, Newark.

(3) Heckman, P.E. "Credibility and Solvency," in Pricing Property and Casualty Insurance Products, Casualty Actuarial Society 116-152 (1980).

(4) Heckman, P.E. and Meyers, G.G. "The Calculation of Aggregate Loss Distributions from Claim Severity and Claim Count Distributions," Proceedings of the Casualty Actuarial Society LXX 111-143 (1983).

(5) Jewell, W.S., "The Use of Collateral Data in Credibility Theory: A Hierarchical Model," Giornale dell' Instituto Italiano degli Attuari, 38, 1-16 (1975)

(6) Patrik, G.S. and John, R.T. "Pricing Excess-of-Loss Casualty Working Cover Reinsurance Treaties," in Pricing Property and Casualty Insurance Products, Casualty Actuarial Society, 399-474 (1980).

(7) Taylor, G.C. "Credibility Analysis of a General Hierarchical Model, Scandanavian Actuarial Journal, 1-12 (1979).