Fitting to Loss Distributions with Emphasis on Rating Variables

Farrokh Guiahi, Ph.D., F.C.A.S, A.S.A

Revised June 2000

Abstract

This paper focuses on issues and methodologies for fitting alternative statistical models--probability distributions--to samples of insurance loss data. The interaction of parametric loss distributions, deductibles, policy limits and rating variables in the context of fitting distributions to losses are discussed. Fitted loss distributions serve an important function for the purpose of pricing insurance products. The procedures illustrated in this paper are based on a sample of insurance losses, and with lognormal as the underlying loss distribution.

Key words

Loss Distributions, Generalized Linear Models, Curve Fitting, Right Censored and Left Truncated data, Rating Variables, Maximum Likelihood Estimation.

1. Introduction

This section presents some preliminaries regarding losses, deductibles, policy limits and rating variables as inputs for fitting distributions to losses. In section 2, a method for fitting a single distribution to losses is considered. In this instance, the information provided by rating variables is either not considered or is not available. The method of maximum likelihood has been applied to estimate model parameters in the presence of deductibles and policy limits. Sections 3 and 4 develop methodologies for fitting alternative statistical models--family of loss distributions--to loss data, using the information provided by rating variables. This is achieved by requiring a parameter of a loss distribution to depend upon values of rating variables. Criteria for assessing goodness of fit are discussed. Furthermore, large sample statistical tests for assessing the impact of rating variables upon loss distributions are given. Some concluding statements are made in section 5.

Insurance data considered here have the following characteristics: a) losses are specified individually, b) for each individual loss, the information about deductibles and policy limits is furnished, and c) for each loss, we have auxiliary policy information regarding the rating variables. Each of these three items is discussed further below.

Losses are given on an individual basis, and have not been grouped by loss size. The methodologies to fit distributions to data differs, depending on whether losses are grouped or individually specified. Losses may be closed or open. The amount

recorded for each loss is the incurred value as of the latest available evaluation period. If some losses in the sample data are still open as of the latest evaluation period, then those losses should be properly adjusted for further development. Unfortunately, most of the methodologies for development of losses to their ultimate values are only available for grouped data. Further research on the topic of development of individual losses to their individual ultimate values is welcomed. Individual losses should be suitably trended to reflect values expected in the future. The methodology presented in this paper has been applied to a sample of commercial fire losses (see Table A of Appendix A). Those losses were mostly closed, as of their latest evaluation date, hence adjustments for further development were not warranted. Finally, in order to fit distributions to losses, zero losses should be excluded.

Deductibles are used to exclude certain losses. Usually deductibles are small--for example, a few hundred or a few thousand dollars. However, for a large insured, deductibles may be sizable due to the existence of self-insured retention or other underlying coverages. Only dollar deductibles are considered here. Time deductibles such as waiting periods are not treated. A reported loss with a value in excess of its deductible is said to be left truncated. If a loss arises from a policy with no underlying deductible, then for the purpose of the computation, a value of zero is imputed as the "deductible" amount. It is not required that the deductible amount be the same for each loss.

Policy limits serve to limit the amount of payment on a given loss or a loss occurrence. When the loss amount is at least as large as its policy limit, the loss is said to have been right censored. If a loss arises from a policy where there is no underlying policy limit, then any amount greater than the loss amount may be imputed as the "policy limit". In these instances, those losses have not been censored. Varying policy limits are allowed for. In fact, no grouping of losses based upon deductible or policy limit amounts is required.

Samples of insurance loss data are usually incomplete. This is due to inclusion of left truncated (losses in excess of deductibles) and right censored (some losses capped by their respective policy limits) data in the sample. Due to this incompleteness of data, it becomes more difficult to estimate the parameters of a loss distribution and to assess the goodness of fit. Many traditional approaches for estimation of parameters of a loss distribution or assessing the goodness of fit of a distribution are valid only if the sample of observations is complete, that is, when there are neither left truncated nor right censored observations in the sample.

Rating variables in insurance depend upon the line of business, the degree of competition present in the market, and regulation. The effect of the rating variables upon loss distributions has important implications for underwriting selection. It also provides for a more differentiated rating system. How to incorporate the information provided by rating variables into the process of fitting distributions to losses is discussed in sections 3 and 4.

Following is a description of how to fit a single distribution to a sample of insurance loss data.

2. Fitting a Single Distribution to Losses

Fitting a single distribution to losses is based upon consideration of alternative statistical models--probability distributions--as data-generating mechanisms. The assumption made is that the observed losses are a realization of a probabilistic process governed by a parametric distribution. The purpose of fitting a distribution to losses is to identify a specific parametric distribution which provides a reasonable fit to the data. A good introduction to the subject of fitting distributions to losses is given by Hogg and Klugman (1984). This paper complements their work by focusing on certain related topics. First, more emphasis is placed on the procedures for fitting loss distributions to individual loss data rather than grouped data. Second, methodologies required to incorporate rating variables in the process of fitting distributions to losses are presented in sections 3 and 4. Finally, readers of this paper may find the computer programs (codes) given here to be beneficial for the purpose of the computing maximum likelihood estimates of parameters of a loss distribution.

Fitting a distribution to losses serves to moderate the effect of sampling variation in the data. This is achieved by replacing an empirical distribution by a more smoothed (fitted)

distribution. Furthermore, estimates of tail probabilities beyond the range of the original data can be provided based on fitted distribution.

At least two problems complicate the fitting of a parametric distribution to loss data. The first problem concerns the tendency of many losses to be settled at rounded figures. This notion is incompatible with selecting a parametric distribution such as lognormal or Pareto, where the probability of taking any specified value is zero. The second problem arises from the fact that many statistical procedures assume that losses in a sample are identically distributed. Insurance risks are normally heterogeneous. Each risk has its own risk characteristics and its own propensity to produce a potential loss. For instance, two different drivers have differing loss propensities. To a certain extent, risk characteristics are reflected by underwriting rating factors. For this reason, risks with the same values for their underwriting factors are cross-classified to produce "homogeneous" classes. The use of rating factors to cope with the heterogeneity problem is addressed in sections 3 and 4. In this section, the information provided by rating factors is ignored in order to concentrate on fitting a single loss distribution to data.

For the sake of exposition, the process of fitting a single distribution to loss data has been broken down into four steps:

1. Consideration of a number of parametric probability distributions as potential candidates for underlying loss distribution.

2. For each distribution specified in step 1, the estimation of the parameters of the distribution from sample data--hence, the determination of a set of fitted distributions.
3. Specification of a criterion for choosing one or a few fitted distributions from step 2 above.
4. Assessing the goodness of fit for the fitted distribution(s) in step 3.

Let us proceed with a more detailed account of these steps. These steps will be illustrated below by reference to a numerical example. The first step requires considering a number of parametric distributions as potential candidates for the data generating mechanism. The list of potential parametric distributions as candidates for loss distribution is enormous. In practice, one can entertain only a few parametric distributions for the purpose of fitting a distribution to losses. In this paper, I have selected the following parametric probability distributions: lognormal, Pareto, Weibull, gamma, inverse gamma, and exponential. This list is subjective, but some of the above distributions have been used by actuaries and have appeared in actuarial literature. The list chosen here is only for illustrative purposes and is not meant to be exhaustive.

The second step involves the estimation of the parameters of each probability distribution selected in step 1 from the data. Once one has estimated the parameters of a given distribution, one then has a fitted distribution. The estimation of parameters of a loss distribution is made difficult because

of incompleteness of data. Some commonly used statistical procedures to estimate parameters of a distribution for a sample of complete data are: the method of moments, the least squares estimation as used for regression models, and the maximum likelihood estimation. These parameter estimation procedures are outlined in most basic statistics texts. For incomplete sample data (presence of left truncated or right censored data), the above estimation procedures are not applicable without further modifications. The application of estimation procedures suitable for complete data to insurance data which is incomplete will produce inefficient parameter estimates. In this paper, the estimation of parameters of a loss distribution is based upon proper specification of the likelihood function reflecting the presence of left truncated and right censored observations in the data.

Following are some necessary notations needed to write an expression for the likelihood function in the case of incomplete data.

Let y_i be the i^{th} loss amount (incurred value), $1 \le i \le n$, where n denotes the number of losses in the data set.

D_i is the deductible for the i^{th} loss.

PL_i is the policy limit for the i^{th} loss.

$f(y;q,j)$ denotes the density function for the loss amount in the case of complete data. q is the primary parameter of interest. j is the nuisance parameter which may be a vector.

$F(y;q,j)$ denotes the cumulative distribution function for the loss amount.

The contribution of a loss to the functional form of the likelihood function depends upon whether the loss is ground-up or in excess of deductible, and furthermore if the loss has been capped by its respective policy limit. Hence, the contribution of a loss to the likelihood function may be one of the four mutually exclusive and exhaustive forms, written as L_{i1}, L_{i2}, L_{i3}, and L_{i4}, as defined below. In addition, four indicator variables, d_{i1}, d_{i2}, d_{i3} and d_{i4} are used in order to write a succinct expression for the likelihood function of the sample.

Case 1: No deductible, and loss below policy limit (neither left truncated nor right censored data). The complete sample case.

$$L_{i1} = f(y_i; q, j) \qquad (2.1a)$$

$$d_{i1} = \begin{cases} 1, \text{If } \mathrm{D_i} = 0 \text{ and } y_i < \mathrm{PL_i} \\ 0, \text{ Otherwise} \end{cases} \qquad (2.1b)$$

Case 2: A deductible, and loss below policy limit (left truncated data)

$$\mathrm{L_{i2}} = \frac{\mathrm{f(D_i + y_i; \theta, \varphi)}}{1 - \mathrm{F(D_i; \theta, \varphi)}} \qquad (2.2a)$$

$$d_{i2} = \begin{cases} 1, \text{If } \mathrm{D_i} > 0 \text{ and } y_i < \mathrm{PL_i} \\ 0, \text{ Otherwise} \end{cases} \qquad (2.2b)$$

Case 3: No deductible, and loss capped by policy limit (right censored data)

$$L_{i3} = 1 - F(\mathrm{PL}_i; q, j) \qquad (2.3a)$$

$$d_{i3} = \begin{cases} 1, \text{If } \mathrm{D_i} = 0 \text{ and } y_i \geq \mathrm{PL_i} \\ 0, \text{ Otherwise} \end{cases} \qquad (2.3b)$$

Case 4: A deductible, and loss capped by policy limit (left truncated and right censored data)

$$L_{i4} = \frac{1 - F(\mathrm{D_i} + \mathrm{PL}_i; q, j)}{1 - F(\mathrm{D_i}; q, j)} \qquad (2.4a)$$

$$d_{i4} = \begin{cases} 1, \text{If } \mathrm{D_i} > 0 \text{ and } y_i \geq \mathrm{PL_i} \\ 0, \text{ Otherwise} \end{cases} \qquad (2.4b)$$

The contribution of the i^{th} loss to the likelihood function is given by

$$L_i = L_{i1}^{d_{i1}} L_{i2}^{d_{i2}} L_{i3}^{d_{i3}} L_{i4}^{d_{i4}} \qquad (2.5)$$

The likelihood function for the sample is given by

$$L = \prod_i L_i \qquad (2.6)$$

The log-likelihood is given by

$$\mathrm{l} = \sum_i \log(\mathrm{L_i}) \qquad (2.7a)$$

$$= \sum_i \mathrm{l}_{ii} \qquad (2.7b)$$

$$\mathrm{l}_i = \log(L_i) \qquad (2.8a)$$

$$= d_{i1} \log(L_{i1}) + d_{i2} \log(L_{i2}) + d_{i3} \log(L_{i3}) + d_{i4} \log(L_{i4}) \qquad (2.8b)$$

where the log, as used in this paper, represents the natural logarithm.

Equation (2.5) represent the contribution of the i^{th} loss to the likelihood function. The likelihood function for the data is given by equation (2.6). To estimate the parameters θ and j we should maximize the likelihood function or alternatively minimize the negative of the logarithm of the likelihood function. Equation (2.7) and (2.8) provide expressions for the logarithm of the likelihood function.

Note that the contribution to the likelihood function for an individual observation in most basic statistics textbooks is of the form (2.1a).

The third step requires a criterion for ranking or comparing alternative fitted probability distributions. This step is needed to reduce the number of fitted distributions in step 2 to one or a few potential candidates. A statistical criterion used for comparing alternative models--statistical distributions--is based upon the value of Akaike's Information Criterion, AIC; refer to Akaike (1973).

The AIC criterion is defined by

AIC = - 2(maximized log-likelihood)
+ 2(number of parameters estimated)

Note, AIC can also be written as

AIC = - 2{maximized log-likelihood – number of parameters estimated}

When two models are compared, the model with a smaller AIC value is the more desirable one.

The AIC is based on log-likelihood and it penalizes the log-likelihood by subtracting for the number of parameters estimated.

Two other model selection criteria used in statistics are Schwarz's Bayesian Information Criterion (BIC), Schwarz(1978), and Deviance as used in Generalized Linear Models; see McCullagh and Nelder (1989). These three criteria are based on maximized log-likelihood function.

Before proceeding to step 4, regarding fit, I shall illustrate steps 1, 2, and 3 by reference to a numerical example. Let us consider the data in Table A of Appendix A.

Here, we have a sample of 100 commercial fire losses. For each loss the deductible, policy limit, and the code for a type of construction are stated. For the time being, let us ignore the information about the construction since we are concerned with fitting a single distribution to the data. For each distribution listed in Table 1 below, I have computed the maximized log-likelihood function, and the corresponding AIC values. For the case of Weibull distribution, the program used to compute the maximum likelihood estimate of parameters and the computed value of maximized log-likelihood function is given as Exhibit 1 in Appendix B. This program is coded in S-Plus, a statistical software suitable for data analysis. The computation of maximized likelihood function for other distributions in Table 1 is similar to the one for Weibull.

Table 1

Distribution	Negative maximized log-likelihood function	Number of Parameters	AIC
lognormal	897.8	2	1799.6
Pareto	895.2	2	1794.4
Weibull	899.8	2	1803.6
gamma	914.5	2	1833.0
inverse gamma	893.7	2	1791.4
exponential	986.4	1	1974.8

With regard to Table 1, it should be noted that the values of maximized likelihood function are positive. The values of logarithm of the maximized likelihood functions are negative and hence the negatives of the logarithm of the maximized likelihood functions are positive figures.

Table 1 can be used for selecting a parametric distribution for the data. Based on the AIC criterion as a method of ranking different fitted distributions, note that the AIC values of lognormal, Pareto, and inverse gamma are "comparable". The AIC values for gamma and exponential distributions suggest relatively more inferior fits. I have selected lognormal, with parameters μ and s^2, as the distribution to be fitted to our data. There are several reasons for this selection. First, it is easier to interpret the parameters of a lognormal distribution. Selecting a simpler model is preferable, as it is easier to explain and comprehend. By taking the logarithm of the losses, the μ parameter represents the location parameter (mean), and the σ parameter is the scale (standard deviation). Second, lognormal distribution has been previously used to describe the distribution of fire losses; see Benckert and Jung (1974).

Now we proceed with step 4, regarding the fit. By examining the data in Appendix A, we note that the losses can be divided into four categories according to four cases defined for specification of the likelihood function (see Table 2 below):

Table 2

#	Case	Number of Losses
1.	No deductible and loss below policy limit	1
2.	A deductible, and loss below policy limit	96
3.	No deductible and loss capped by policy limit	0
4.	A deductible and loss capped by policy limit	3

For our data, most of the losses are of case 2, i.e., losses with deductibles and values below their policy limits. Due to the paucity of data, we concentrate only on case 2.

For lognormal distribution, we can compute theoretical conditional distributions (probabilities) and conditional limited expected values based on a fitted distribution, and compares these quantities with their respective sample counterparts.

The conditional distribution or probability of a lognormal random variable, X, with parameters μ and σ is given by

$$P(X \le b \mid X > a) = \frac{\Phi(\frac{\log(b)-m}{s}) - \Phi(\frac{\log(a)-m}{s})}{1-\Phi(\frac{\log(a)-m}{s})}$$

where Φ is the cumulative distribution function of a standard normal distribution. Here a represents a threshold or a deductible amount D, and b is usually the sum of deductible and limit, i.e., D + PL.

The conditional limited expected value is defined by

$$E[\min(X,b) \mid X > a] = \frac{1}{1-\Phi(\frac{\log(a)-m}{s})} \{ e^{m+\frac{1}{2}s^2} [\Phi(\frac{\log(b)-m-s^2}{s}) - \Phi(\frac{\log(a)-m-s^2}{s})] + b[1-\Phi(\frac{\log(b)-m}{s})] \}$$

Table 3 summarizes the comparison of theoretical and sample values of conditional probabilities and conditional limited expected values for case 2 of data in Appendix A.

Table 3

Comparison of Conditional Probabilities and Conditional Limited Expected Value for Fitted Lognormal with its Sample Values

a = 500

b	$P(X \le b \mid X > a)$ Based on lognormal*	$P(X \le b \mid X > a)$ Sample estimate	$E[\min(X,b) \mid X > a]$ Based on lognormal*	$E[\min(X,b) \mid X > a]$ Sample estimate
2,000	0.485	0.494	1,538.7	1,620.9
5,000	0.714	0.699	2,666.4	2,737.2
10,000	0.832	0.843	3,747.2	3,764.3
20,000	0.909	0.904	4,969.3	4,907.7
30,000	0.938	0.952	5,716.8	5,547.9
40,000	0.954	0.976	6,248.3	5,833.6
50,000	0.964	0.988	6,655.8	6,071.7

* $\hat{m}$ = 5.887, $\hat{s}$ = 2.302 are the maximum likelihood estimates for the fitted lognormal distribution.

The comparisons of fitted and sample quantities in Table 3 suggests the lognormal provided a "reasonable" fit to the data. It is worth making a few comments regarding fit. First, our sample size is 100, with 96 observations for case 2. With small sample sizes, considerable sampling variability are encountered in estimation of model parameters. Second, a perfect fit implies no smoothing! Third, the fit for a specific type of distribution is judged to be good if it has a high predictive power, that is, whether the same type of distribution provides good fits to many samples of the same kind. A quotation from Lindsey (1995), is appropriate here: "*If a model represents the sample too well, it will have no chance of representing a second, similarly generated, sample very well. A model too close to a sample will usually be too far from the population.*" Finally, it is worth emphasizing that there are many other possible potential candidates (probability distributions) for fitting to a specific data set. Thus, curve fitting is to some extent subjective and not a perfect science. From a practical point of view, there are other considerations related to fitting a distribution to a sample. These are: a) the volume and quality of data, b) the time constraint in which to do the curve fitting, c) the knowledge and experience of the curve fitter, d) availability of suitable software (programs), e) convergence of iterative algorithms for estimation of model parameters, and specification of initial values for parameters, and f) the treatment of outliers. Last but not the least is consideration of the purpose

for which the fitted distribution is used. With all these qualifications regarding fit, we shall assume the lognormal provides a reasonable fit to the data in Appendix A.

3. Fitting a Family of Distributions to Loss Data: A Mean Approach

In section 2, procedures to fit a single distribution to loss data were considered. The information provided by rating variables was not considered. As mentioned earlier, risks in insurance tend to be heterogeneous. Risks with different attributes may well have different loss distributions. To a certain degree, a risk's characteristics are reflected through the values pertained by its rating variables. Thus, we expect the loss distribution for fire for a small unprotected frame building be different from a large, highly protected and fire-resistive building. It is desirable to have loss distributions which reflect these differences. Our approach to this issue is to construct suitable statistical models--family of loss distributions. Two possible solutions are proposed in this paper. The first solution, as explained in this section, is similar in spirit to the Generalized Linear Models (GLM) approach. An excellent account on the subject of GLM is given by McCullagh and Nelder (1989). An alternative solution is presented in section 4.

Loss distributions dependent upon rating variables have important implications for underwriting selection and

determination of rates. By including the rating variables, one generally improves the fit to the data. Using statistical models enables one to assess the effect of rating variables on loss distributions by performing statistical tests of hypotheses.

A traditional approach for obtaining loss distributions dependent upon risk attributes is to segment losses into subgroups. Then, for each subgroup, a separate fitted loss distribution is obtained. For instance, in fire insurance, losses may be classified broadly by construction as frame, masonry and fire-resistive. Three fitted loss distributions can be obtained according to the types of construction. Segmentation of data into classes gives rise to credibility problems. For the problem alluded to, it would be exasperating if one considered eight construction types instead of three, and in addition, considered other rating factors such as protection and occupancy.

In section 2, we noted that the lognormal distribution provides a reasonable fit to the data in Appendix A. Mirroring the approached used in GLM, let us now fit a family of lognormal distributions to our data.

The GLM methodologies consist of three components. These are referred to as the random component, the systematic component, and the link. The random component: the random variable of interest, Y (e.g., losses) or a transformation of Y, has a distribution belonging to the exponential family of distributions. The density, in canonical form, for the exponential family is

$$f(y;q,j) = \exp\{[(qy - b(q))/a(j)] + c(y,j)\}$$

where a(.), b(.) and c(.) are some specific functions. q is the primary parameter of interest, and j is often referred to as the nuisance parameter. Suitable loss distributions in the exponential family include normal, gamma and inverse Gaussian.

The systematic component of a GLM specifies the explanatory variables, $x_1, x_2, \ldots, x_p$ (e.g., rating variables). The explanatory variables may only influence the distribution of the Y through a single linear function called the linear predictor η,

$$h = b_0 + b_1 x_1 + \ldots + b_p x_p$$

The link, g, specifies how the mean of Y, E(Y), is related to the linear predictor η, i.e.

$$g(E(Y)) = h = \sum_j b_j x_j$$

The form of the link function varies by the type of distribution within the exponential family of distributions. For the normal distribution the link function is the identity map, i.e., μ = η.

In GLM, the information provided by explanatory variables (rating variables) is summarized by a linear predictor. Each explanatory variable is considered either as a factor (categorical) or as a covariate (quantitative). For instance, sex, construction, and protection are categorical in nature, while age and amount of insurance are quantitative.

Some additional notations are needed to specify our statistical model. Let h_i denote a linear predictor for the i^{th}

loss. It summarizes the information conveyed by the rating variables for the ith loss. We write

$$h_i = x_i^T \boldsymbol{b}$$

$$= \sum_{j=0}^{p} x_{ij} b_j$$

$$= b_0 + \sum_{j=1}^{p} b_j x_{ij}$$

where β is a $(p+1) \times 1$ vector of unknown parameters. x_i is a $(p+1) \times 1$ vector of known constant terms, x_{ij}'s. The first element of x_i, x_{i0} is set equal to one. Its purpose is to represent a constant term (intercept) in the expression for the linear predictor. The other x_{ij}'s components, $1 \le j \le p$, are used to represent rating variables. The value of p is partially dependent upon the number of categorical rating factors included in the model, as well as their respective number of levels (values). In addition, p depends upon the number of quantitative rating variables in the model. Note that when rating variables are not taken into consideration, or the information about them is not available, then p takes on the value of zero. This corresponds to the fitting of a single distribution to the entire loss data as described in section 2.

Following are some examples of the linear predictors, h_i, to be discussed throughout this paper. Some commonly used categorical rating factors in fire insurance are construction, protection, and occupancy. The amount of insurance (insured building value), a measure of exposure, is quantitative. Here,

we shall consider only construction and building value for illustrative purposes. Assume there are three possible construction types (levels), namely frame, masonry and fire-resistive. In GLM, as well as regression analysis, the contribution of a categorical variable to a linear predictor is made by specifying dummy variables. For the construction rating factor, we need to introduce two dummy variables C_{i1} and C_{i2}, defined as follows:

$$C_{i1} = \begin{cases} 1, \text{If the i}^{\text{th}} \text{ risk is a frame} \\ 0, \text{Otherwise} \end{cases}$$

$$C_{i2} = \begin{cases} 1, \text{If the i}^{\text{th}} \text{ risk is a masonry} \\ 0, \text{Otherwise} \end{cases}$$

For the i^{th} loss, let $\mathrm{BV_i}$ denote the amount of insurance purchased by the policyholder to cover damages arising from peril of fire to the building. For a fire policy, the policy limit for the building cover is synonymous with the building value. Since there is a wide range of variability among building values, we shall use the logarithm of the building value instead of building value as our covariate in the linear predictor. For these two variables, namely, construction and building value, we shall define four statistical models corresponding to four linear predictors as follows:

Model A: $h_i = b_0$ (3.1A)

Model B: $h_i = b_0 + b_1 C_{i1} + b_2 C_{i2}$ (3.1B)

Model C: $\eta_i = \beta_0 + \beta_1 \log(\mathrm{BV_i})$ (3.1C)

Model D: $h_i = b_0 + b_1 \log(\mathrm{BV_i}) + b_2 C_{i1} + b_3 C_{i2}$ (3.1D)

The linear predictor given by equation (3.1A) is used when either we do not take into consideration the information given by rating variables or when no information on rating variables is available. In these instances, we are fitting a single distribution to the entire data. We shall refer to this Model A as the "base" model (distribution). The base distribution is used as a benchmark to gauge the relative improvement in fit by including rating variables.

The linear predictor corresponding to (3.1B) is appropriate if construction is the only rating factor used. Using the statistical methodology developed here, the entire data is used to estimate the values of the parameters b_0, b_1, b_2 simultaneously. This approach is different from the one in which the data is segmented into three sub-groups according to types of construction.

The linear predictor (3.1C) is used when we wish to examine only the effect of exposure size (building value) on loss distribution.

Finally, we shall use (3.1D) when both construction and building value are considered. In this case, the vector $x_i^T = (1 \;\; \log(\mathrm{BV_i}) \;\; C_{i1} \;\; C_{i2})$ represents the contribution of the i^{th} risk's attributes to the linear predictor, and p has the value of three.

The four linear predictors given by (3.1A), (3.1B), (3.1C), and (3.1D) generate four statistical models. This is an example of nested models. For nested models, some models are a special case of a more general model. The linear predictors

(3.1A), (3.1B) and (3.1C) are special cases of the linear predictor (3.1D). For the linear predictor (3.1D), Model D, we can entertain the following statistical tests of hypotheses:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \qquad (3.2)$$

$$H_0: \beta_2 = \beta_3 = 0 \qquad (3.3)$$

$$H_0: \beta_1 = 0 \qquad (3.4)$$

The null hypothesis (3.2) is used to test if either construction or building value (exposure size) has any effect on loss distribution. The acceptance of this null hypothesis, subject to the usual interpretation of Type Two error probability, suggests that the rating variables have no appreciable influence on the loss distribution. The rejection of (3.2) implies that the inclusion of building value or construction in the linear predictor gives a superior model as compared to the fit by the base distribution, Model A. The acceptance of the null hypothesis (3.3) suggests that in the presence of building value, the addition of the construction factor does not improve the fit. Null hypothesis (3.4) can be similarly interpreted.

By conducting statistical tests corresponding to the previously stated hypotheses, the effects of rating variables on loss distributions can be assessed. The test statistics are likelihood ratio tests. The asymptotic distribution of test statistics are Chi-squares. Hence, for small sample sizes, the implications of the above tests based on Chi-squares are only approximately valid.

Here, we assume that the underlying loss random variable, Y_i--for the i^{th} risk--has a lognormal distribution with

parameters μ_i and s^2. The parameter μ_i is the mean of transformed variable log(Y_i). We shall refer to models in this section as "Mean" models. Using an approach similar to GLM, we relate the rating variables of interest to parameter μ_i by using an identity link function. That is,

$$m_i = x_i^T b$$

$$= b_0 + \sum_i x_{ij} b_j$$

where $b_0, b_1, \ldots, b_p$ are regression like parameters and x_{ij}'s represent the contribution of explanatory rating variables for the i^{th} risk. Hence, we have a family of lognormal distributions, with parameters $b_0, b_1, \ldots, b_p$ and s^2 to describe the distribution of losses.

It is assumed that the parameter σ is the same for each risk, and does not vary by the rating variables. We shall examine an alternative approach in the next section, where σ is not constant. Although, the mean and variance of the loss distributions vary by rating variables, but due to the constancy of σ, the skewness, and the kurtosis are not dependent on rating variables.

The mechanism to fit a family of lognormal distributions to the data of Table A of Appendix A has now been established. A set of nested hypotheses of interest, (3.2), (3.3), and (3.4) in reference to model (3.1D) has also been stated. We now need to perform the necessary computations to estimate the model parameters, and calculate log-likelihood statistics for

alternative models as described by linear predictors (3.1A), (3.1B), (3.1C), and (3.1D).

The program to compute maximum likelihood estimate of model parameters for the linear predictor (3.1D), as well as the value of the negative of log-likelihood based upon maximum likelihood estimates is given as Exhibit 2 of Appendix B.

Likelihood ratio test statistics are needed for performing nested statistical tests of hypothesis (3.2), (3.3), and (3.4). The likelihood ratio test statistics can be calculated from the values of log-likelihood statistics for the appropriate models.

The upper portion of Table 4 below provides the values of the negative of log-likelihood statistics for the “mean” models according to linear predictors (3.1A), (3.1B), (3.1C), and (3.1D). The lower portion of Table 4, provides the values of the necessary likelihood ratio test statistics for performing nested statistical hypotheses (3.2), (3,3), and (3.4). In addition, the appropriate 95^{th} percentiles and degrees of freedom of the asymptotic distributions of test statistics are also provided.

Table 4

Likelihood Statistics for Alternative Statistical Models

"Mean" Models

Model	Linear Predictor	Negative of logarithm of Likelihood function
A	$m_i = b_0$	897.7654
B	$m_i = b_0 + b_1 C_{i1} + b_2 C_{i2}$	894.8344
C	$m_i = b_0 + b_1 \log(BV_i)$	896.8284
D	$m_i = b_0 + b_1 \log(BV_i) + b_2 C_{i1} + b_3 C_{i2}$	892.7099

Nested Hypotheses based on Model D

Test of Hypothesis	Likelihood Ratio* Test Statistics	DF for Chi-sq.	95th perc. of Chi-sq.
$H_0: \beta_1 = \beta_2 = \beta_3 = 0$	$-2(\log L_A - \log L_D) = 10.1110$	3	7.8147
$H_0: \beta_2 = \beta_3 = 0$	$-2(\log L_C - \log L_D) = 8.2370$	2	5.9915
$H_0: \beta_1 = 0$	$-2(\log L_B - \log L_D) = 4.2490$	1	3.8415

* L_A, L_B, L_C, and L_D, above, correspond to likelihood statistics for "Mean" Models A, B, C, and D respectively.

Let us interpret the results given by Table 4, later on we shall make some qualifications regarding our interpretations.

If we are interested to test whether construction factor or building value has an effect upon the shape of the loss distribution, the appropriate null hypothesis is $H_0: \beta_1 = \beta_2 = \beta_3 = 0$. The value of the test statistic, i.e., the likelihood ratio test statistics is 10.111. Since 10.111 exceeds the value of 7.8147 (the boundary of rejection region), it implies that we should reject the null hypothesis H_0. The implication is either construction or building value have an influence on the shape of the loss distribution. Similar interpretations can be given for the other two null hypotheses.

Some qualifications regarding the above interpretation of Table 4 are in order. First, due to relatively small sample size, and the approximate distribution of likelihood ratio test, as Chi-squares, we should be careful to interpret the results given in Table 4. Second, the numerical estimate of parameters (see Exhibit 2 of Appendix B) and the implications of the nested test of hypotheses, are only for illustrative purposes and are not intended to be used for any rating purposes.

Finally, the Model D has the largest likelihood value. Based upon the values of likelihood statistics, as well as the AIC values, Model D fits the data better than Model A, the base distribution. Recall that Model A corresponds to the case of fitting a single distribution to the data. Thus, the consideration of rating variables has led to an improvement in fit, and this improvement is statistically significant.

4. Fitting a family of Lognormal Distributions with Different Scale Parameters

In section 3, a family of lognormal distributions using a procedure “similar” to the GLM approach was introduced. These alternative statistical models were referenced to as "Mean" models. The linear predictor was set equal the μ parameter of the lognormal, and the σ parameter was assumed to be constant. By considering the logarithm of losses, log(Y), the rating variables affected the mean of the distribution but not the scale, the σ parameter. In this section, a family of lognormal distributions is introduced where the scale σ is made to depend on rating variables, and the parameter μ is treated as a constant. Using methodology similar to that in section 3, four new statistical models A, B, C, and D, are defined corresponding to four linear predictors as follows:

Model A: $s_i = b_0$ (4.1A)

Model B: $s_i = b_0 + b_1 C_{i1} + b_2 C_{i2}$ (4.1B)

Model C: $s_i = b_0 + b_1 \log(\mathrm{BV_i})$ (4.1C)

Model D: $s_i = b_0 + b_1 \log(\mathrm{BV_i}) + b_2 C_{i1} + b_3 C_{i2}$ (4.1D)

These models will be referred to as "Scale" models. Parallel to the development in section 3, we have three nested statistical hypotheses of interest for Model D, linear predictor (4.1D), defined as

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \qquad (4.2)$$

$$H_0: \beta_2 = \beta_3 = 0 \qquad (4.3)$$

$$H_0: \beta_1 = 0 \qquad (4.4)$$

The purpose and interpretation of these hypotheses is similar to those of (3.2), (3.3), and (3.4) of section 3.

With the mechanism established in section 3, we want to evaluate the it of alternative "Scale" models fitted to the data in Table A of Appendix A. The results of these computations are summarized in Table 5 below. A program for the maximum likelihood estimate of parameters, and likelihood statistics for Model B, linear predictor (4.1B), is given in Exhibit 3 of Appendix B. For comparison purposes, the values of likelihood ratio statistics for the "Mean" models are also reproduced in Table 5.

Table 5

Likelihood Statistics for Alternative Statistical Models

"Scale" Models

Model	Linear Predictor	Negative of logarithm of Likelihood function
A	$s_i = b_0$	897.7654
B	$s_i = b_0 + b_1 C_{i1} + b_2 C_{i2}$	892.4242
C	$s_i = b_0 + b_1 \log(\mathrm{BV}_i)$	895.7967
D	$s_i = b_0 + b_1 \log(\mathrm{BV}_i) + b_2 C_{i1} + b_3 C_{i2}$	887.9109

Nested Hypotheses Based On Model D Comparison of "Mean" & "Scale" Models

Test of Hypothesis	Likelihood Ratio Test Statistics*	Mean Model	Scale Model	DF for Chi-sq.	95th perc. of Chi-sq.
$H_0: b_1 = b_2 = b_3 = 0$	$-2(\log L_A - \log L_D)$	10.1110	19.7090	3	7.8147
$H_0: \beta_2 = \beta_3 = 0$	$-2(\log L_C - \log L_D)$	8.2370	15.7716	2	5.9915
$H_0: \beta_1 = 0$	$-2(\log L_B - \log L_D)$	4.2490	9.0266	1	3.8415

*Depending upon the context, the L_A, L_B, L_C, and L_D, above, correspond to likelihood functions for "Mean" or "Scale" Models A, B, C, and D.

Once again we should be careful to interpret the results given in Table 5 due to relatively small sample size, and the approximate distribution of likelihood ratio test as Chi-squares. With these qualifications in mind, it appears that the "Scale" models provide a better fit than the "Mean" models to our data.

5. Conclusion

This paper discusses issues related to curve fitting. It provides appropriate statistical methodologies for fitting parametric distributions to loss data. In particular, the interaction of parametric probability distributions, deductibles, policy limits and rating variables are considered. The presence of deductibles and policy limits complicate the estimation of parameters of loss distribution, and the assessment of goodness of fit. Procedures to fit a single distribution or a family of distributions to loss data were given. Statistical tests of hypotheses to assess the effect of rating variables upon loss distribution were discussed. The methodologies developed in this paper were applied to a sample of loss data using lognormal as the reference distribution. Sample programs coded in S-Plus, a statistical package, were provided to illustrate the numerical computation of maximum likelihood estimate of model parameters and maximized likelihood function. Finally, the results in this paper suggest that for any specific data set, there may be many viable statistical

models suitable for the purpose of fitting distributions to the data.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., *Second International Symposium on Inference Theory*, Budapest: Akademiai Kiado, pp. 267-281.

Benckert, L.G. and Jung, J. (1974). Statistical Models of Claim Distributions in Fire Insurance. *ASTIN Bulletin* ***8****, 1-25.*

Hogg, R.V. and Klugman, S.A. (1984). *Loss Distributions.* John Wiley & Sons, New York.

Lindsey, J. K. (1995). *Introductory Statistics: A Modelling Approach*. Oxford University Press.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Models*, Second Edition. Chapman and Hall, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.

Appendix A: TABLE A

Deductible	Policy Limit	Loss	Construction	Deductible	Policy Limit	Loss	Construction
1,000	57,000	502	2	250	43,000	75	2
250	41,000	31,971	1	1,000	1,000	865	3
1,000	1,000	367	1	100	33,000	206	2
250	60,000	698	2	250	7,000	2,303	1
100	10,000	4,863	2	250	64,000	11,760	2
250	24,000	834	2	250	45,000	402	2
250	16,000	646	1	500	30,000	3,352	1
250	60,000	198	2	250	2,000	511	1
1,000	66,000	275	2	0	10,000	1,115	2
250	36,000	500	1	250	52,000	237	2
100	53,000	1,518	2	250	3,000	1,197	2
250	70,000	2,430	2	100	50,000	7,107	2
250	51,000	357	1	250	89,000	535	2
250	79,000	2,008	2	1,000	200,000	5,959	2
500	139,000	3,044	1	250	100,000	1,224	3
250	155,000	238	2	250	85,000	85,000*	1
250	150,000	3,244	2	250	103,000	2,358	2
250	98,000	850	2	250	110,000	31,243	2
250	100,000	198	2	500	110,000	1,488	1
100	110,000	110,000*	1	250	175,000	2,702	3
250	115,000	1,191	1	1,000	154,000	850	2
250	100,000	1,852	3	250	100,000	300	2
5,000	153,000	4,433	1	250	134,000	930	2
250	120,000	100	2	500	125,000	305	2
250	100,000	2,501	2	1,000	115,000	190	2
250	350,000	1,057	2	250	630,000	1,875	1
250	373,000	180	1	1,000	402,000	5,075	2
1,000	208,000	9,385	1	500	204,000	972	2
1,000	600,000	2,300	3	250	300,000	271	3
1,000	284,000	5,589	1	250	350,000	87	1
1,000	263,000	652	2	500	595,000	625	2
250	312,000	3,975	1	1,000	275,000	20,934	1
250	280,000	485	2	250	290,000	609	1
1,000	312,000	2,092	2	250	560,000	325	2
2,500	250,000	250,000*	1	1,000	371,000	6,012	1
250	300,000	250	2	1,000	362,000	860	2
500	625,000	1,305	3	250	317,000	2,720	2
1,000	319,000	6,729	3	500	6,817,000	1,040	3
500	9,214,000	185	2	1,000	3,010,000	48,762	1
1,000	3,000,000	22,930	3	5,000	6,023,000	20,576	1
1,000	800,000	498	3	250	700,000	230	2
500	838,000	990	2	1,000	1,000,000	200	2
250	1,400,000	5,491	3	500	1,442,000	1,247	1
1,000	1,500,000	1,185	3	1,000	2,000,000	10,000	2
500	36,819,000	6,032	2	1,000	2,526,000	4,525	3
250	1,282,000	13,775	2	500	65,065,000	16,981	2
250	1,000,000	150	3	1,000	1,236,000	4,911	2
1,000	6,127,000	4,536	2	1,000	5,000,000	81,692	2
100	1,140,000	298	3	250	2,275,000	21,447	2
1,000	1,910,000	335	2	1,000	2,700,000	992	2

*Building losses with asterisks next to them are losses capped by their respective insured building values (right censored.)

Appendix B: Exhibit 1

An S-Plus Program to Compute Maximum Likelihood Estimate of Parameters & Maximized Likelihood Statistic for Weibull Distribution

```
mydata<-TableA
m<-data.frame(mydata)
Weibull<-function(lamda, alfa, data = data.matrix)
  {    D <- data.matrix[,1]
      PL <- data.matrix[,2]
       y <- data.matrix[,3]
      z <- D+((y <PL)*y+(y >=PL)*PL)
    delta1<- (D==0)*(y <PL)
    delta2<- (D> 0)*(y <PL)
    delta3<- (D==0)*(y >=PL)
    delta4<- (D> 0)*(y >=PL)
    L1 <- alfa*lamda*(z^(alfa-1))*exp(-lamda*(z^alfa))
    L2 <-(alfa*lamda*(z^(alfa-1))*exp(-lamda*(z^alfa)))/exp(-lamda*(D^alfa))
    L3 <- exp( - lamda * (z^alfa))
    L4 <- exp( - lamda * (z^alfa))/exp( - lamda * (D^alfa))
    logL<- delta1*log(L1)+delta2*log(L2)+delta3*log(L3)+delta4*log(L4)
   -logL  }
min.Weibull<-ms(~Weibull(lamda,alfa), data=m, start
=list(lamda=1,alfa=.15))
min.Weibull
value: 899.802
parameters:
     lamda     alfa
 0.4484192 0.223073
formula:  ~  Weibull(lamda, alfa)
100 observations
call: ms(formula =  ~ Weibull(lamda, alfa), data = m, start = list(lamda
= 1, alfa = 0.15))
```

S-Plus is a statistical package produced by StatSci, a division of MathSoft, Inc., Seattle, Washington.

Weibull density is: $f(x;\lambda,\alpha) = \alpha\ \lambda\ x^{\alpha-1}\exp(-\lambda\ x^{\alpha})$

Appendix B: Exhibit 2

An S-Plus Program to Compute Maximum Likelihood Estimate of Parameters & Maximized Likelihood Statistic for a Family of Lognormal Distributions Based on “Mean” Model D

```
mydata<-TableA
m<-data.frame(mydata)
lognormal.model.D <- function(b0,b1,b2,b3,sigma, data=data.matrix)
  {  D <- data.matrix[,1]
    PL <- data.matrix[,2]
     y <- data.matrix[,3]
     z <- D+(y*(y<PL)+PL*(y>=PL))
  cnst <- data.matrix[,4]
    C1 <- cnst == 1
    C2 <- cnst == 2
     d <-D+(D == 0)*1
    mu <- b0+b1*log(PL)+b2*C1+b3*C2
      delta1 <- (D == 0)*(y < PL)
      delta2 <- (D > 0)*(y < PL)
      delta3 <- (D == 0)*(y >= PL)
      delta4 <- (D > 0)*(y >= PL)
      L1 <- dlnorm(z,mu,sigma)
      L2 <- dlnorm(z,mu,sigma)/(1-plnorm(d,mu,sigma))
      L3 <- 1-plnorm(z,mu,sigma)
      L4 <- (1-plnorm(z,mu,sigma))/(1-plnorm(d,mu,sigma))
      logL <-delta1*log(L1)+delta2*log(L2)+delta3*log(L3)+delta4*log(L4)
 -logL  }
min.model.D<-ms(~lognormal.model.D(b0,b1,b2,b3,sigma), data=m,
 start=list(b0=4.568, b1=0.238, b2=1.068, b3=0.0403, sigma=1.322))
min.model.D
value: 892.7099
parameters:
       b0        b1       b2        b3    sigma
 1.715296 0.3317345 2.154994 0.4105021 1.898501
formula:  ~  lognormal.model.D(b0, b1, b2, b3, sigma)
100 observations
call: ms(formula = ~ lognormal.model.D(b0, b1, b2, b3, sigma), data=m,
start =list(b0=4.568, b1=0.238, b2=1.068, b3=0.0403, sigma=1.322))
```

Appendix B: Exhibit 3

An S-Plus Program To Compute Maximum Likelihood Estimate of Parameters & Maximized Likelihood Statistic for a Family of Lognormal Distributions Based on "Scale" Model B

```
mydata<-TableA
m<- data.frame(mydata)
lognormal.Scale.model.B<- function(b0,b1,b2,mu, data=data.matrix)
   {  D <- data.matrix[,1]
      PL <- data.matrix[,2]
      y <- data.matrix[,3]
    cnst <- data.matrix[,4]
      z <- D + (y*(y < PL)+PL*(y >= PL))
      C1 <- cnst == 1
      C2 <- cnst == 2
      d <- D + (D == 0) * 1
      sigma <- b0+b1*C1+ b2* C2
      delta1 <- (D == 0)*(y < PL)
      delta2 <- (D > 0)*(y < PL)
      delta3 <- (D == 0)*(y >= PL)
      delta4 <- (D > 0)*(y >= PL)
      L1 <- dlnorm(z,mu,sigma)
      L2 <- dlnorm(z,mu,sigma)/(1 - plnorm(d,mu,sigma))
      L3 <- 1 - plnorm(z,mu,sigma)
      L4 <- (1 - plnorm(z,mu,sigma))/(1 - plnorm(d,mu,sigma))
      logL <-delta1*log(L1)+delta2*log(L2)+delta3*log(L3)+delta4*log(L4)
   -logL   }
 min.Scale.B<- ms(~lognormal.Scale.model.B(b0,b1,b2,mu), data=m,
+ start=list(b0=2,b1=0,b2=0,mu=6))
 min.Scale.B
value: 892.4242
parameters:
       b0       b1        b2      mu
 1.583642 1.324647 0.1066956 6.55098
formula:  ~  lognormal.Scale.model.B(b0, b1, b2, mu)
100 observations
call: ms(formula =  ~ lognormal.Scale.model.B(b0, b1, b2, mu), data = m,
start = list(b0 = 2, b1 = 0, b2 = 0, mu = 6))
```