Modelling Operational Losses

Ancus Roehr

The Advanced Measurement Approach (Basel Committee on Banking Supervision 2001) for operational risk capital calculation requires banks to collect operational loss data. Data pools are forming in the industry and external vendors are offering database subscriptions to assist banks in acquiring sufficient data. This article addresses the question of how to set up an actuarial loss model with such heterogeneous data. Special consideration is given to sets of losses in excess of various thresholds and how they might be combined to calibrate a single model. The sensitivity of model parameters to those varying thresholds is analysed.

Since the Basel Committee (the Committee) presented its proposal for a New Capital Accord, banks are facing a new type of regulatory capital charge: one for operational risk. Various schemes for calculating that charge have been proposed by the Committee (cf. Basel Committee on Banking Supervision 2001), ranging from a simplistic Basic Indicator Approach to Advanced Measurement Approach(es) that apply sophisticated quantification techniques. Banks using the latter can hope to reduce their capital charge in comparison to the Standardized Approach by up to 25%. To do so, they must demonstrate to their regulator that their models are robust, verifiable, and prove that a lesser charge will suffice. The set up of these models, has not been, and probably will not be, specified in detail by the Committee. However, it is proposed that the operational risk capital charge be the Unexpected Loss at a confidence level of 99.9%.

Another proposed cornerstone of the process is an internal loss database, with at least five years of data in it (three years during a transition period), plus the evaluation of data available from external sources. These sources can be data pools that banks form among themselves, such as the now defunct MORE consortium, the effort undertaken by the BBA or the new ORX consortium. Others are external loss databases such as Zurich IC^2's F1RST.

Such data is particularly useful in conjunction with what is called the Loss Distribution Approach (LDA) or an actuarial model by the Basel Committee and what is known as frequency/severity models in insurance. The approach is justified, among other reasons, by the fact that operational risk losses are largely (stochastically) independent, that is, correlation is spurious and not a dominant feature. Furthermore, parametric severity models are able to extrapolate the tail of the distribution into high regions where empirical data is scarce. This is important because it is the tail of the distribution that dominates the calculation of the Unexpected Loss.

If the tail dominates the calculation, it seems natural to set up the severity model only in excess of some relatively high loss threshold. This approach is also supported by a result from Extreme Value Theory, which says that Generalized Pareto distributions are good approximations to tails in excess of high thresholds under very general assumptions.

Thresholds enter the picture also on the data collection side—most data are collected with some threshold in force, so that losses below that threshold are not recorded in the database. This measure applies to internal databases, data pools and external databases alike. In general, thresholds vary depending on the institution collecting the data. For example, the recent Second Quantitative Impact Study carried out by the Basel Committee used a threshold of EUR 10,000 in their survey. The data in the F1RST database, by contrast, are generally collected at a threshold of USD $1,000,000.

Hence, combining internal and external data leads to mixed sets of data—mixed in the sense that different subsets of the data have been collected at different thresholds. But, even for the internal data of a specific bank, the threshold may change over time and vary across business lines or loss types.

Outline and summary of paper

The purpose of this paper is to analyse how to fit actuarial models to mixed sets of loss data. First, maximum likelihood estimation is used in this setting. It turns out that one can no longer separate the estimation of the frequency from that of the severity. Next, the pros and cons of fitting analytic severity curves piece-wise along the severity axis, as suggested frequently by practitioners, are discussed briefly. Typical usage is to fit the tail in excess of a high threshold using one curve, and fit another curve in the region below that threshold. The second curve may fit to zero, or only to a second, lower, threshold. It may be an empirical distribution, given enough data. The results herein show that empirical distributions for the lower part will be of little use in conjunction with mixed sets of data, or where the high threshold is so high that few empirical data points exceed it.

The sensitivity of loss severity model parameter estimation with respect to the threshold of a data sample is discussed. Specifically, suppose a parametric loss severity model is to be fitted in excess of a threshold z_0, and a sample of data has been collected at a (higher) threshold z_1, we investigate how confidence intervals for the model parameters depend on the threshold z_1. In the examples, the investigation is carried out for some of the most widely used loss severity distributions, namely, lognormal, (Generalized) Pareto and Weibull. The results show that, in most cases, and with the size of the sample kept fixed, confidence intervals widen as the threshold moves up; but we also identify cases where they remain constant or even become narrower.

Some remarks on notation

Loss sizes are nonnegative real numbers typically denoted by x or x_i. Loss thresholds are denoted using the variables z or z_1. The probability density of a loss size distribution is denoted $f(x)$, the corresponding distribution function $F(x)$ and the tail function $\bar{F}(x) = 1 - F(x)$. In considering parametric families of distributions, the letter θ is used to denote the parameter, or the vector of parameters, as in $f(x, \theta), F(x, \theta)$, while θ_i denotes one of the components of θ. Certain standard regularity assumptions, such as differentiability and integrability, are usually not stated explicitly.

Estimating parameters with mixed data

When collecting losses, one usually discards losses below a certain threshold z so as to make the task manageable and cost-efficient. This threshold, however, may change over time depending on the line of business or loss type. Data shared with data pool members and external databases are likely to come with different thresholds. In the case of external databases, the thresholds will likely be much higher than in any internal database.

To complicate matters further, just as losses have to be scaled to compensate, for example, for inflation, so must thresholds. It is therefore advisable

for all data collection efforts to store the collection threshold with each loss record.

The goal, now, is to fit an actuarial model to a mixture of samples collected at various thresholds using the method of maximum likelihood estimation.

Suppose one has m loss data sets collected at thresholds $z_1, \ldots, z_m$. Let S_i be the set corresponding to the threshold z_i, and let $x_{i1}, \ldots, x_{in_i}$ be the losses in that set. Note that $x_{ij} \geq z_i$. Assume that all losses are stochastically independent throughout the remainder of this article.

Suppose the losses from set S_i have been collected over a period of t_i years. These losses may come from a data pool representing a number of banks, or they represent more or less the whole banking industry, as in the case of external databases. To account for this, a volume factor w_i is introduced.

In the simplest case, when the base of the data sample is a certain group of banks whose characteristics (e.g., size, organization, business and the like) are the same as those of the one under consideration, the factor w_i is just the number of members of that group. In other cases, determination of an appropriate w_i is a difficult scaling problem, comparable to the scaling of loss sizes.

Note that the z_i do not have to be pairwise different. For example, the first sample might consist of n_1 losses collected internally $(w_1 = 1)$ at threshold $z_i = \text{EUR } 5000$ over $t_1 = 2.5$ years, while the second sample of size n_2 may come from a data pool, collected at the same threshold $z_2 = \text{EUR } 5000$ over $t_2 = 4$ years, and the estimate may be that the pool data gets a weight of $w_2 = 5$.

The loss model consists of a loss size density $f(x,\theta)$ with support on the interval $[z_0, \infty)$, where $z_0 \leq z_i$, $i = 1, \ldots, m$ and a Poisson distribution with mean λ for the annual frequency of loss. Recall that the Poisson parameter λ coincides with the mean and the variance of the Poisson distribution. The model is applied to a time horizon of one year, and so the time factor is $t = 1$. Assigning the bank in this model a "unit" size, the volume factor is $w = 1$. Noting that t and w are linked as a factor in many instances, a variable $v = tw$ is also defined.

Since the sum of two independent Poisson experiments is again a Poisson experiment, the Poisson parameter is proportional to time and volume, and one would expect λv_i losses, on average, in set S_i. In other words, one expects n_i, the number of losses observed in S_i, to be a realization of a Poisson experiment with Poisson parameter λv_i. However, this holds true only if S_i had been collected at the model threshold z_0. Only a fraction of losses exceed the higher threshold z_i. That fraction, according to the model, is given by the tail probability $\bar{F}(z_i, \theta)$. Hence, the model predicts $\lambda_i := \lambda t_i w_i \bar{F}(z_i,\theta)$ losses, on average, in S_i, and by a well-known property of the Poisson distribution, the frequency n_i of losses in excess of z_i is again Poisson distributed (hence, it has a Poisson parameter λ_i). Furthermore, each (independent) loss in S_i is a realization of a random variable with density $\frac{f(x,\theta)}{1-F(z_i,\theta)}$.

Note that this is the conditional (on the threshold) severity distribution of losses in the model.

The likelihood for the outcome n_i of a Poisson experiment with parameter λ_i is

$$e^{-\lambda_i}\frac{\lambda_i^{n_i}}{n_i!}$$

and given n_i independent losses $x_{i1},\ldots,x_{in_i}$ in S_i, their combined likelihood is

$$\prod_{j=1}^{n_i}\frac{f(x_{ij},\theta)}{1-F(z_i,\theta)}.$$

As we assume n_i and $x_{i1},\ldots,x_{in_i}$ to be independent random variables, the total likelihood of S_i is

$$e^{-\lambda_i}\frac{\lambda_i^{n_i}}{n_i!}\prod_{j=1}^{n_i}\frac{f(x_{ij},\theta)}{1-F(z_i,\theta)}.$$

For all our (independent) samples combined, we arrive at a likelihood function of

$$L := \prod_{i=1}^{m} e^{-\lambda_i}\frac{\lambda_i^{n_i}}{n_i!}\prod_{j=1}^{n_i}\frac{f(x_{ij},\theta)}{1-F(z_i,\theta)}$$

which is equal to

$$\prod_{i=1}^{m}\left(e^{-\lambda v_i \bar{F}(z_i,\theta)}\frac{(\lambda v_i)^{n_i}}{n_i!}\prod_{j=1}^{n_i} f(x_{ij},\theta)\right).$$

Applying the logarithm and dropping terms independent of the model parameters, the problem translates to the maximization of

$$\sum_{i=1}^{m}\left(-\lambda v_i \bar{F}(z_i,\theta) + n_i \log\lambda + \sum_{j=1}^{n_i}\log f(x_{ij},\theta)\right).$$

This can be attacked in the usual manner, by finding a value of (λ,θ) where all partial derivatives with respect to λ and the θ_k vanish.

The derivatives with respect to λ and θ_k, respectively, are

$$\sum_{i=1}^{m}\left(-v_i \bar{F}(z_i,\theta) + \frac{n_i}{\lambda}\right),$$

and

$$\sum_{i=1}^{m}\left(-\lambda v_i \partial_{\theta_k}\bar{F}(z_i,\theta) + \sum_{j=1}^{n_i}\partial_{\theta_k}\log f(x_{ij},\theta)\right).$$

The former vanishes for

$$\hat{\lambda} = \frac{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} v_i \bar{F}(z_i,\theta)}, \qquad (1)$$

which leaves only the equation

$$\partial_{\theta_k}\log\sum_{i=1}^{m} v_i \bar{F}(z_i,\theta) = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\partial_{\theta_k}\log f(x_{ij},\theta) \qquad (2)$$

to be solved for θ, where

$$n := \sum_{i=1}^{m} n_i.$$

If time or volume factors (hence the v_i) are not available for all samples, or if one is interested in a loss size model only, one may use the likelihood function

$$L = \prod_{i=1}^{m}\prod_{j=1}^{n_i}\frac{f(x_{ij},\theta)}{1-\bar{F}(x_{ij},\theta)},$$

which does not involve the parameter λ. Taking logarithms and derivatives, this leads to equations

$$\sum_{i=1}^{m} n_i \partial_{\theta_k}\log \bar{F}(z_i,\theta) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\partial_{\theta_k}\log f(x_{ij},\theta) \qquad (3)$$

to be solved for θ. The result may then be inserted into Equation 1, where only those samples for which v_i exists are used in the formula.

In the special case $z_i = z_0$, for all i, $\bar{F}(z_i,\theta) = 1$ for all i, and Equation 1 is just the total number of losses in all samples, divided by the sum of the weights. In particular, estimators for λ and θ can be derived independently of each other. In the more general case of mixed thresholds, this is not true.

As an example, consider the Pareto case with parameter $\alpha > 0$. In this situation,

$$\bar{F}(x,\alpha) = \left(\frac{x}{z_0}\right)^{-\alpha},$$

and

$$f(x,\alpha) = \frac{\alpha}{z_0}\left(\frac{x}{z_0}\right)^{-\alpha-1}.$$

Since

$$\partial_\alpha \log\bar{F}(x,\alpha) = -\log\frac{x}{z_0},$$

and

$$\partial_\alpha \log f(x,\alpha) = \frac{1}{\alpha} - \log\frac{x}{z_0},$$

Equation 3 becomes

$$\sum_{i=1}^{m} -n_i \log\frac{z_i}{z_0} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\frac{1}{\alpha} - \log\left(\frac{x_{ij}}{z_0}\right)\right).$$

The solution is

$$\hat{\alpha} = \frac{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} \log\,(x_{ij}/z_i)} \qquad (4)$$

at which point L attains a maximum since $\partial_\alpha \partial_\alpha \log L = -\alpha^{-2} n < 0$. Equation 2, on the other hand, cannot be solved for α explicitly.

Fitting tails

One option is to split the loss severity model into two regions: one in excess of a high threshold z_1, and the other below that threshold, say in the interval $[z_0, z_1]$, where z_0 is another, smaller threshold (which may be zero). For example, if none of the standard parametric loss severity models show a good fit to the data on the interval $x > z_0$, then splitting the region and fitting models separately may help. One may even want to use an empirical loss distribution in the lower region. Another excellent reason to fit the tail separately comes from Extreme Value Theory. A famous result from that theory says that under fairly general conditions, tails can be approximated by Generalized Pareto distributions, at least in excess of high thresholds. An example illustrates this point later on.

Fitting a tail beyond the high threshold requires data in that region. Such data may not be available from an internal loss database, but data pools and external databases can be very useful here. However, external data will most likely only serve the purpose of fitting the *conditional* distribution of losses exceeding z_1. Gluing that distribution to the lower region requires knowledge of the total probability mass of the tail in excess of z_1. It may be possible to obtain a non-parametric estimate from internal loss data, but only if some of the internal losses—hopefully more than just a very few—exceed the threshold z_1. In many instances this will not be the case.

In the lower region, empirical loss models may not be so simple to set up if one is working with mixed sets of loss data that attach at various thresholds in between z_0 and z_1. One solution is to discard all sets with thresholds greater than z_0, but that may also mean ignoring much valuable data. To resolve this problem, fit a parametric model—as described above—to the mixed data in the lower region. From this model one can also readily obtain an estimate for the tail mass in excess of z_1, which solves the other problem mentioned above.

Sensitivity of parameters

If we want to set up a loss severity model in excess of the threshold z_0, but have a sample of data that has been collected at a higher threshold $z > z_0$, how much information will that sample reveal about the model parameters? Are sometimes large losses from external databases particularly helpful in estimating "tail parameters"? These questions are examined in the context of maximum likelihood estimation, but restricted to the case where only the severity model, not the frequency model, is to be calibrated.

In the Pareto case, the only parameter is the parameter α, which completely determines the shape and, in particular, the tail of the distribution. Looking at Equation 4 for the maximum likelihood estimator α, note first that each summand x_{ij}/z_i in the denominator, viewed as a random variable, follows the same distribution: a Pareto distribution on $[1, \infty)$ with tail $x^{-\alpha}$. Exactly which sample S_i it belongs to is unimportant; regardless of the threshold, each loss has the same potential impact on the estimator for α. In this sense, samples above high thresholds are neither better nor worse than samples (of equal size) above a low threshold to determine the distribution parameter. Luckily, in this case, one can write down the maximum likelihood estimator in closed form. However, this is rarely the case.

The general case is more difficult to analyse. In particular, the maximum likelihood estimator cannot be written in closed form. Fortunately, the Fisher Information is a good gauge of the information content of the data sample with respect to answering questions about the underlying parameters θ. For a single observation with underlying probability density function $f(x,\theta)$, the Fisher Information is defined as the matrix $I = \left(I_{\theta_j \theta_k}\right)$ of expected values

$$I_{\theta_j\theta_k}(\theta)$$
$$= E((\partial_{\theta_j}\log f(x,\theta))(\partial_{\theta_k}\log f(x,\theta)))$$
$$= \int_0^\infty (\partial_{\theta_j}\log f(x,\theta))(\partial_{\theta_k}\log f(x,\theta))f(x,\theta)dx$$

provided the integrals exist. Integration is normally over the support of the distribution, here the support is assumed to be $[0,\infty)$.

The Fisher Information matrix is defined for each observation, with f being the probability density underlying that observation. Here, sets of samples that attach at various thresholds are of interest, so f has to be adapted to each threshold. The total Fisher Information of the sample then is just the sum of the individual Fisher Information values (Rao 1965, 5a.4).

If $g = g(\theta)$ is a function of the model parameters, and $\nabla_g = (\ldots, \partial_{\theta_j} g, \ldots)$ its vector of partial derivatives with respect to the model parameters, then the Cramér-Rao inequality (Rao 1965, 5a.3) states that for any unbiased estimator $\hat{g}$ of g, its variance

$$\mathrm{Var}(\hat{g})$$

satisfies

$$\mathrm{Var}(\hat{g}) \geq (\nabla g) I^{-1} (\nabla g)$$

and so is bounded from below. Although unbiased estimators for which the inequality is sharp do not always exist, one may still use the right-hand side of this inequality as a gauge of the level of uncertainty about the parameters that is to be expected in a given situation.

Note that if $\boldsymbol{I}$ is the Fisher Information of a sample of size 1, then a sample of size n has the Fisher Information $n\boldsymbol{I}$, which reduces the bound for the inequality by a factor of n compared to the sample size 1.

Recall that the elements of a sample above a threshold z follow a distribution with density

$$\frac{f(x,\theta)}{1-F(z,\theta)}.$$

Hence, the dependence of the Fisher Information matrix on z is as follows:

$$I_{\theta_j\theta_k}(z,\theta)$$
$$= \int_z^\infty \left(\partial_{\theta_j}\log \frac{f(x,\theta)}{1-F(x,\theta)}\right)$$
$$\cdot\left(\partial_{\theta_k}\log \frac{f(x,\theta)}{1-F(x,\theta)}\right)\frac{f(x,\theta)}{1-F(z,\theta)}dx.$$

Using

$$\int_z^\infty \partial_{\theta_j} f(x,\theta)dx = \partial_{\theta_j}\int_z^\infty f(x,\theta)dx$$
$$= \partial_{\theta_j}\bar{F}(z,\theta),$$

(which assumed that differentiation and integration commute), and suppressing the dependence on θ, this can be rewritten as

$$I_{\theta_j\theta_k}(z) = -(\partial_{\theta_j}\log \bar{F}(z))(\partial_{\theta_k}\log \bar{F}(z))$$
$$+\frac{1}{\bar{F}(z)}\int_z^\infty (\partial_{\theta_j}\log f(x))$$
$$\cdot(\partial_{\theta_k}\log f(x))\, f(x)dx. \qquad (5)$$

As shown in Appendix A, under certain regularity conditions, which are satisfied in the later examples, this can also be expressed as

$$I_{\theta_j\theta_k}(z) = \frac{1}{\bar{F}(z)}\int_z^\infty G_{\theta_j\theta_k}(x)\, f(x)dx$$
$$= \frac{\int_z^\infty G_{\theta_j\theta_k}(x)\, f(x)dx,}{\int_z^\infty f(x)dx}$$

where

$$G_{\theta_j\theta_k}(x,\theta) = G_{\theta_j}(x,\theta)G_{\theta_k}(x,\theta),$$

and

$$G_{\theta_j}(x,\theta) = \partial_{\theta_j}\log\frac{f(x,\theta)}{1-F(x,\theta)}$$
$$= \partial_{\theta_j}\log(-\partial_x \log \bar{F}(x,\theta)).$$

With this machinery in place, the following examples examine the Fisher Information and derived variance bounds as a function of the thresholds for a variety of common distributions.

Examples

The examples presented here are the Pareto, Generalized Pareto, lognormal and Weibull distributions. The Pareto distribution is analytically tractable and conservative, in that it is too heavy tailed in most cases; it can only be used for tail-fitting data. This also applies to the Generalized Pareto distribution, but it is a two-parameter family—one more than the Pareto distribution—and, hence, offers more flexibility. Its use is also justified by a result from Extreme Value Theory (see the Generalized Pareto example below). The lognormal distribution often fits loss data quite well across the whole positive axis. The Weibull distribution is less often used, but it is presented here because it exhibits some unusual properties.

Pareto

In the Pareto case,

$$I_{\alpha\alpha}(z,\alpha) = \frac{1}{\alpha^2}$$

is independent of the threshold z. This is easily computed, as shown in Appendix B. It is also in line with earlier observations.

Generalized Pareto

The Generalized Pareto distribution is given by the tail

$$\bar{F}(x,\xi,\beta) := \left(1+\xi\frac{x}{\beta}\right)^{-\frac{1}{\xi}}, \quad if\ \xi>0$$

$$\bar{F}(x,\xi,\beta) := e^{-\frac{x}{\beta}}, \quad if\ \xi=0$$

with scaling parameter $\beta>0$. The support of the distribution is on $x\geq 0$. (The first definition makes sense also for negative ξ, but then the distribution has finite support on $0\leq x/\beta\leq -\xi^{-1}$. We will not consider this case in the sequel.)

One of the main results in Extreme Value Theory states that for a large class of distributions, tails can be ("asymptotically") approximated by distributions of the above type. For details, cf. Embrechts et al. (1997), Theorem 3.4.14. This result has tended to lead practitioners to fit Generalized Pareto distributions to data right away, although the result just mentioned holds only in the limit where the threshold tends to infinity and approximation errors may be significant.

The Fisher Information matrix is easy to calculate in the case $\xi = 0$, because here we can refer to Appendix A from which it follows that $I(z) = \beta^{-2}$ independent of z.

If $\xi>0$, then one obtains $I(z)^{-1} = (1+\xi)\mathbf{M}$ where the matrix $\mathbf{M}$ contains the elements

$$\mathbf{M}_{11} = (1+\xi)$$

$$\mathbf{M}_{12} = -\beta\left(1+(1+2\xi)\frac{z}{\beta}\right)$$

$$\mathbf{M}_{21} = -\beta\left(1+(1+2\xi)\frac{z}{\beta}\right)$$

$$\mathbf{M}_{22} = \beta^2\left(2+2(1+2\xi)\frac{z}{\beta}+(1+\xi)(1+2\xi)\left(\frac{z}{\beta}\right)^2\right).$$

$\mathbf{M}_{11}$ being independent of z, we see that, similar to the Pareto case, the influence on ξ does not depend on the threshold, whereas the influence on estimating β diminishes as z grows, because $\mathbf{M}_{22}$ is an increasing function of z.

Lognormal

In the lognormal case,

$$f(x,\mu,\sigma) = \frac{e^{-\frac{1}{2}\left(\frac{\log x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\,x\,\sigma},$$

with μ and σ the mean and standard deviation, respectively, of the logarithm of the lognormal random variable. Letting $\bar{\phi}$ be the tail function of the standard normal distribution and setting

$$u(x) = \frac{\log x-\mu}{\sigma},$$

one obtains

$$\bar{F}(x) = \bar{\phi}(u(x)),$$

and, with the abbreviation

$$h(u) = \frac{e^{\frac{-u^2}{2}}}{\sqrt{2\pi}},$$

(the density function of the standard normal distribution) one verifies

$$\partial_\mu f(x) = \frac{uf(x)}{\sigma},$$
$$\partial_\sigma f(x) = \frac{(u^2-1)f(x)}{\sigma},$$
$$\partial_\mu \bar{F}(x) = \frac{h(u)}{\sigma},$$
$$\partial_\sigma \bar{F}(x) = \frac{uh(u)}{\sigma}.$$

Note that

$$\int_z^\infty g(u(x))\, f(x)dx = \int_{u(z)}^\infty g(u)\, h(u)du,$$

and that $\int_{u(z)}^\infty t^n\, h(t)dt$ equals

(as functions of $u = u(z)$)

$$\bar{\phi},\ h,\ \bar{\phi}+uh,\ (2+u^2)h,\ 3\bar{\phi}+(3+u^2)uh$$

for $n = 0, 1, 2, 3, 4$, respectively. One is now well-equipped to calculate $\mathbf{I}$, arriving at

$$\boldsymbol{I}(z) = \frac{1}{\sigma^2}\begin{pmatrix} 1-H(H-u) & H(1-u(H-u)) \\ H(1-u(H-u)) & 2+Hu(1-u(H-u)) \end{pmatrix},$$

with inverse

$$\boldsymbol{I}(z)^{-1} = \frac{\sigma^2}{2+H(H-u)(u(H-u)-3)} \cdot \begin{pmatrix} 2+Hu(u(H-u)-1) & H(u(H-u)-1) \\ H(u(H-u)-1) & 1-H(H-u) \end{pmatrix},$$

where $H = H(u) = \frac{h(u)}{\bar{\phi}(u)}$.

Further, note that

$$\boldsymbol{I}(0) = \frac{1}{\sigma^2}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

To examine how the Cramér-Rao lower bound for the variance of estimators of μ and σ, respectively, varies as the threshold moves up from zero (corresponding to $u = -\infty$), the two diagonal elements of $\boldsymbol{I}^{-1}$ as a function of u are plotted relative to their values at threshold zero. (See Figure 1.)

Notice that if the threshold is the median of the underlying distribution—corresponding to $u = 0$—the variance bound for estimators of μ is already more than 22 times that at threshold zero. Whereas, for estimators of σ, it is only just over eight times as big as at zero. The gap widens as the threshold increases. This result is in line with the intuition that samples at high thresholds should not be of great relevance in determining μ, the logarithm of the median of the distribution.

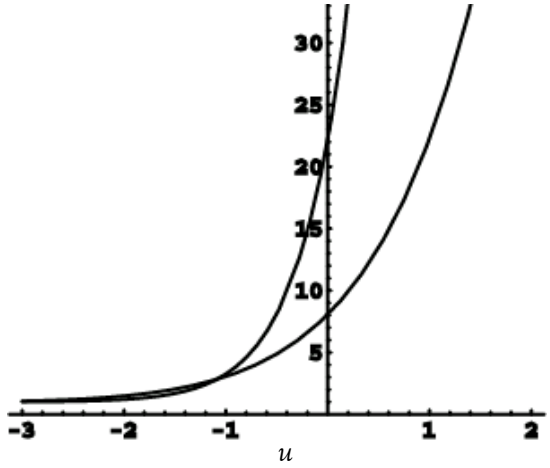

Figure 1: The Cramér-Rao lower bound as a function of u

So, for lognormal models, samples beyond a threshold lose their power in estimating model parameters as the threshold increases. Note that samples of equal size are compared in making this statement.

A slightly different picture emerges if one considers mixed samples, say one big sample of internal losses and another smaller one from an external source. To fix ideas, assume that the internal losses have been collected at threshold 0, the external losses at z, and that the size of the external sample is ε times the sizes of the internal one, with ε a small positive number. The combined Fisher Information is then $\boldsymbol{I}(0) + \varepsilon\boldsymbol{I}(z)$ (times the size of the internal sample). The Cramér Rao inequality becomes, dropping the sample-size factor,

$$\begin{aligned} \mathrm{Var}(\hat{g}) &\geq (\nabla g)(\boldsymbol{I}(0)+\varepsilon\boldsymbol{I}(z))^{-1}(\nabla g) \\ &= (\nabla g)\boldsymbol{I}(0)^{-1}(\nabla g) \\ &\quad -\varepsilon(\nabla g)\boldsymbol{I}(0)^{-1}\boldsymbol{I}(z)\boldsymbol{I}(0)^{-1}(\nabla g) + O(\varepsilon^2). \end{aligned}$$

The change in the bound for the estimators of μ and σ is proportional, respectively, to the first and second diagonal entries of $\mathbf{I}(z)$, because $\mathbf{I}(0)$ is a diagonal matrix. But, while the first entry is a decreasing function of the threshold, the second as a function of u, and in units of σ^{-2}, looks as follows. (See Figure 2.)

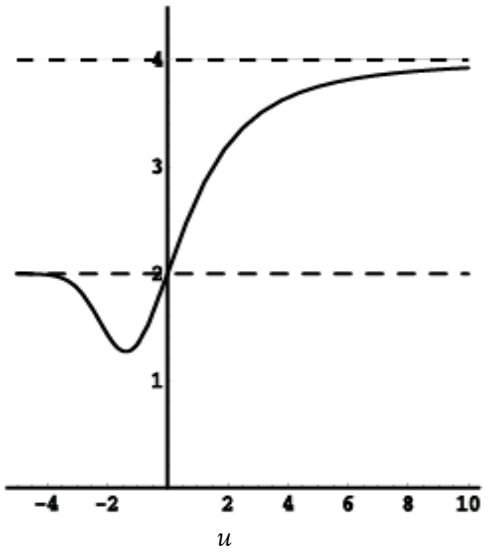

Figure 2: The second diagonal entry of the inverse of the Fisher Information matrix as a function of u

So at $z = 0$ (equivalent to $u = -\infty$), $I_{\sigma\sigma}(z) = 2\sigma^{-2}$ and the same value is attained at the median (equivalent to $u = 0$). Beyond the median, $I_{\sigma\sigma}(z)$ increases and approaches a value twice as high as at threshold zero. Hence, a small sample at a high threshold brings the variance bound for estimators of σ down more than one of equal size at a lower threshold.

One-parameter Weibull

The one-parameter family with tail function

$$\bar{F}(x,a) = e^{-x^a},$$

where $a > 0$ is a family of Weibull-type distributions on the nonnegative real axis $x \geq 0$.

Typically, the Weibull distribution is given with two parameters. Here, one of these, the scaling parameter, has been dropped. It can be re-introduced by replacing x with x/b on the right-hand side of the above equation.

One computes

$$G_{aa}(z,a) = \left(\frac{1}{a} + \log(z)\right)^2,$$

and so, by the result in Appendix A, $I_{aa}(z,a)$ is eventually growing to infinity as z increases.

This is an example of a distribution where losses in excess of some high threshold reveal more about the underlying parameter a than the same number of losses collected in excess of some low threshold. This behaviour is the exception rather than the rule.

Conclusions

When fitting frequency/severity models to several sets of loss data collected at different thresholds, the most natural approach is to estimate frequency and severity parameters jointly within the framework of maximum likelihood estimation. However, this approach requires volume information, the v_i factors that may not be available for all sets of data, especially from external databases. In that case, one may still apply maximum likelihood estimation for the severity part only.

External data, usually at high thresholds, can be particularly valuable in estimating some of the severity parameters, as is shown by our analysis of the Fisher Information. Incidentally, it is a simple matter to extend that analysis to the case of joint estimation of severity and frequency parameters.

References

Basel Committee on Banking Supervision, 2001, "Working Paper on the Regulatory Treatment of Operational Risk," Basel: Bank For International Settlements.

Embrechts, P., C. Klüppelberg and T. Mikosch, 1997, *Modelling External Events for Insurance and Finance*, 2nd Printing, 1999, Berlin: Springer.

Rao, C. R., 1965, *Linear Statistical Inference and Its Applications*, 2nd ed., 1973, New York, NY: John Wiley & Sons.

Appendix A: The Fisher Information as a function of the threshold

We will prove the following in this section:

Let

$$G_{\theta_j}(x,\theta) := \partial_{\theta_j} \log \frac{f(x,\theta)}{1-F(x,\theta)} = \partial_{\theta_j} \log\left(-\partial_x \log \bar{F}(x,\theta)\right)$$

and

$$G_{\theta_j\theta_k} := G_{\theta_j} G_{\theta_k}.$$

Then the Fisher Information matrix $\boldsymbol{I} = \boldsymbol{I}(z,\theta)$ satisfies the differential equation

$$(\partial_z \boldsymbol{I})\bar{F} = f(\boldsymbol{I} - \boldsymbol{G})$$

with solution

$$\boldsymbol{I}(z,\theta) = \frac{\boldsymbol{I}(0,\theta) - \int_0^z \boldsymbol{G}(x,\theta)\, f(x,\theta)\,dx}{\bar{F}(x,\theta)},$$

and if

$$\lim_{x\to\infty} \bar{F}(x,\theta)(\partial_{\theta_j} \log \bar{F}(x,\theta))(\partial_{\theta_k} \log \bar{F}(x,\theta)) = 0,$$

then

$$I_{\theta_j\theta_k}(z,\theta) = \frac{\int_z^\infty G_{\theta_i\theta_k}(x,\theta)\, f(x,\theta)\,dx}{\int_z^\infty f(x,\theta)\,dx},$$

and

$$\lim_{z\to\infty} I_{\theta_j\theta_k}(z,\theta) = \lim_{z\to\infty} G_{\theta_j}(z,\theta) G_{\theta_k}(z,\theta),$$

if either limit exists in $\boldsymbol{R}\cup\{\pm\infty\}$.

Proof: Note that for any function $g(z,x)$

$$\partial_z \int_z^\infty g(z,x)\,dx = -g(z,z) + \int_z^\infty \partial_z g(z,x)\,dx$$

provided the integrals exist and differentiation and integration commute. Now, with

$$\partial_z \frac{f(x,\theta)}{\bar{F}(z,\theta)} = \frac{f(x,\theta)}{\bar{F}(z,\theta)} \frac{f(z,\theta)}{\bar{F}(z,\theta)},$$

and

$$\partial_z \partial_{\theta_j} \log \frac{f(x,\theta)}{\bar{F}(z,\theta)} = \partial_{\theta_j} \frac{f(z,\theta)}{\bar{F}(z,\theta)},$$

we see that

$$\partial_z I_{\theta_j\theta_k}(z,\theta) = -G_{\theta_j\theta_k}(z,\theta)\frac{f(z,\theta)}{\bar{F}(z,\theta)} + \frac{f(z,\theta)}{\bar{F}(z,\theta)} \int_z^\infty \left(\partial_{\theta_j} \log \frac{f(x,\theta)}{\bar{F}(z,\theta)}\right) \cdot \left(\partial_{\theta_k} \log \frac{f(x,\theta)}{\bar{F}(z,\theta)}\right) \frac{f(x,\theta)}{\bar{F}(z,\theta)}\,dx,$$

since

$$\begin{aligned}
&\int_z^\infty \left(\partial_z \left(\partial_{\theta_j} \log \frac{f(x,\theta)}{\bar{F}(z,\theta)}\right)\right)\left(\partial_{\theta_k} \log \frac{f(x,\theta)}{\bar{F}(z,\theta)}\right)\frac{f(x,\theta)}{\bar{F}(z,\theta)}\,dx \\
&= \left(\partial_{\theta_j} \frac{f(z,\theta)}{\bar{F}(z,\theta)}\right)\int_z^\infty \left(\partial_{\theta_k} \log \frac{f(x,\theta)}{\bar{F}(z,\theta)}\right)\frac{f(x,\theta)}{\bar{F}(z,\theta)}\,dx \\
&= \left(\partial_{\theta_j} \frac{f(z,\theta)}{\bar{F}(z,\theta)}\right)\int_z^\infty \left(\partial_{\theta_k} \frac{f(x,\theta)}{\bar{F}(z,\theta)}\right)\,dx \\
&= \left(\partial_{\theta_j} \frac{f(z,\theta)}{\bar{F}(z,\theta)}\right)\left(\partial_{\theta_k} \int_z^\infty \left(\frac{f(x,\theta)}{\bar{F}(z,\theta)}\right)\,dx\right) \\
&= 0.
\end{aligned}$$

Hence, the differential equation for $\boldsymbol{I}$. Using $f = -\partial_z \bar{F}$ again, we deduce from it

$$(\partial_z \boldsymbol{I})\bar{F} + \boldsymbol{I}(\partial_z \bar{F}) = \partial_z(\boldsymbol{I}\bar{F}) = -\boldsymbol{G}f$$

and, since $\bar{F}(0,\theta) = 1$, we arrive at

$$\boldsymbol{I}(z,\theta)\bar{F}(z,\theta) = \boldsymbol{I}(0,\theta) - \int_0^z \boldsymbol{G}(x,\theta)\, f(x,\theta)\,dx.$$

But, we saw above that

$$I_{\theta_j\theta_k}(z)\bar{F}(z) = -\bar{F}(z)(\partial_{\theta_j} \log \bar{F}(z))(\partial_{\theta_k} \log \bar{F}(z)) + \int_z^\infty (\partial_{\theta_j} \log f(x)) \cdot (\partial_{\theta_k} \log f(x)) f(x)\,dx,$$

where we have suppressed the dependence on θ to simplify notation.

Subtracting these two equations, we get

$$\begin{aligned}
0 = &\int_0^z (\partial_{\theta_j} \log f(x))\,(\partial_{\theta_k} \log f(x)) f(x)\,dx \\
&- \int_0^z G_{\theta_j\theta_k}(x,\theta)\, f(x,\theta)\,dx \\
&+ \bar{F}(z)(\partial_{\theta_j} \log \bar{F}(z))(\partial_{\theta_k} \log \bar{F}(z)),
\end{aligned}$$

and with our assumption that the last term converges to zero as z goes to infinity, we obtain

$$I_{\theta_j\theta_k}(0,\theta) = \int_z^\infty G_{\theta_j\theta_k}(x,\theta)\, f(x,\theta)dx$$

and, therefore,

$$I_{\theta_j\theta_k}(z,\theta) = \frac{1}{\bar{F}(z,\theta)}\int_z^\infty G_{\theta_j\theta_k}(x,\theta)\, f(x,\theta)dx$$
$$= \frac{\int_z^\infty G_{\theta_j\theta_k}(x,\theta)\, f(x,\theta)dx}{\int_z^\infty f(x,\theta)dx}.$$

Finally, the statement about the limit follows from de L'Hospital's Rule.

Q.E.D.

Appendix B: Fisher Information independent of the threshold

Looking at the differential equation, namely, $(\partial_z I)\bar{F} = f(\boldsymbol{I}-\mathbf{G})$ from Appendix A, we see that $\partial_z \boldsymbol{I} \equiv 0$ implies $\boldsymbol{I} \equiv \mathbf{G}$ and vice versa. The latter implies that $\boldsymbol{I}$ has rank 1, since $\mathbf{G}$ has rank 1. We therefore restrict our attention to one-parameter families of distributions.

If the tail of a one-parameter family of distributions is of the form

$$\bar{F}(x,\theta) = r(x)^{-s(\theta)},$$

then

$$G_\theta(x,\theta) = \left(\partial_\theta \log\left(-\partial_x \log\left(r(x)^{-s(\theta)}\right)\right)\right)^2$$
$$= \left(\partial_\theta \log\left(s(\theta)\right)\right)^2$$

is independent of x (and one can show that the converse holds, too). In this case, $\boldsymbol{I} = \mathbf{G}$ and the information content is independent of the threshold. Examples of one-parameter distributions of this kind are the Pareto ($r(x) = x$, $s(\theta) = \theta$) and the exponential ($r(x) = e^x$, $s(\theta) = \theta$) distributions. If, further, $s(\theta) = e^\theta$, then we even have $\mathbf{G} = 1$, independent also of the underlying parameter.

This latter case deserves some special attention, since it allows one to easily calculate exact confidence intervals for estimators of the model parameter. So, let us consider the family

$$\bar{F}(x,\theta) = r(x)^{-e^\theta}, \qquad x \geq z_0,$$

where r has to be a monotonically increasing function with $r(z_0) = 1$. Both the Pareto and the exponential family can be written in this form by a change of coordinate for the parameter. If a prime denotes differentiation with respect to x, the density is

$$f(x,\theta) = e^\theta r'(x) r(x)^{-e^\theta - 1}$$

and maximum likelihood estimation applied to the losses $x_1,\ldots,x_n$ gives the estimator

$$\hat{\theta} = \log n - \log \sum_{k=1}^{n} \log\, r(x_k).$$

The distribution of each x_k is given by the density f. Applying the monotonous and increasing function $g := \log \circ r$ to the random variable x_k, the density of the new random variable is

$$\frac{f(h(w))}{g'(h(w))}$$

as a function of $w \geq 0$, where h is the inverse of g. We have dropped the dependence on θ here to simplify notation. Let r^{-1} denote the inverse function of r. Then, $h(w) = r^{-1}(e^w)$, and we obtain as density

$$\frac{e^\theta r'(h(w)) e^{-we^\theta - w}}{\frac{r'(h(w))}{r(h(w))}} = e^\theta e^{-e^\theta w},$$

which is the density of an exponentially distributed random variable. Going back to our maximum likelihood estimator $\hat{\theta}$, the sum of the logarithms is a sum of i.i.d. random variables, all exponentially distributed with parameter e^θ. Hence, the sum is $\Gamma(e^\theta, n)$-distributed with density

$$\frac{e^\theta}{(n-1)!}\,(e^\theta x)^{n-1} e^{-e^\theta x}, \quad x \geq 0.$$

The logarithm of the sum has density

$$\frac{e^{(y+\theta)n}}{(n-1)!e^{y+\theta}}, \quad y \in (-\infty,\infty)$$

with mean $-\theta + \psi^{(0)}(n)$ and variance $\psi^{(1)}(n)$, the ("polygamma") functions $\psi^{(0)}$ and $\psi^{(1)}$ being the first and second logarithmic derivatives of the gamma function, respectively. The following formulae hold:

$$\psi^{(0)}(n) = -\gamma + \sum_{k=1}^{n-1} \frac{1}{k},$$

$$\psi^{(1)}(n) = \sum_{k=n}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} - \sum_{k=1}^{n-1} \frac{1}{k^2},$$

where $\gamma = 0.577216\ldots$ is the Euler-Mascheroni constant.

Therefore, if we replace our maximum likelihood estimator $\hat{\theta}$ with the estimator

$$\bar{\theta} := \psi^{(0)}(n) - \log \sum_{k=1}^{n} \log r(x_k),$$

then this estimator is unbiased and has density

$$\frac{e^{-\bar{y}n}}{(n-1)!e^{e^{-\bar{y}}}},$$

where $\bar{y} = y - \theta + \psi^{(0)}(n)$. It is easy to find the integral of this density; let

$$P(j,x) := \sum_{k=j}^{\infty} \frac{x^k}{k!},$$

be the terms of order j and higher in the power series of e^x about $x = 0$, then the tail of the distribution is

$$\int_{\bar{y}}^{\infty} \frac{e^{-sn}}{(n-1)!e^{e^{-s}}}\, ds = \frac{P(n,e^{-\bar{y}})}{P(0,e^{-\bar{y}})} = \frac{P(n,e^{-\bar{y}})}{e^{-\bar{y}}}.$$

Summarizing, we see that the shape of the distribution of $\bar{\theta}$—and in particular the variance—depends only on n and not on the true and, of course, unknown underlying parameter θ, which merely has the effect of a shift. The variance $\psi^{(1)}(n)$ is asymptotically equal to $1/n$, the Cramér-Rao lower bound (remember the Fisher Information is 1 in this parameterization, see above). Finally, with the explicit formulae derived, it is easy to find exact confidence intervals for the unknown parameter without using an asymptotic approximation.