RC23676 (W0507-172) July 25, 2005
Mathematics

IBM Research Report

Modeling of Risk Losses Based on Incomplete Data

Emmanuel Yashchin
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. Ithas been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distributionoutside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g, payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at http://domino.watson.ibm.com/library/CyberDig.nsf/home

Modeling of Risk Losses Based on Incomplete Data

Emmanuel Yashchin Mathematical Sciences Department
IBM Research Division
T.J. Watson Research Center
Yorktown Heights, NY 10598

Key words and phrases: Extreme values, Incomplete data, Operational Risk, Pareto distribution, Size-biased sampling

Abstract

In this article we present a method for drawing inference about the process of losses experienced in relation with operations of a business. For example, for a bank such losses could be related to erroneous transactions, human error, fraud, harassment lawsuits or a power outage. Information about frequency and magnitude of losses is obtained through search of a number of sources, such as printed or Internet based publications related to Insurance and Finance. The data consists of losses that were discovered in the search; it is assumed that the probability of a loss to appear in the body of sources and be discovered increases with its magnitude. Our approach is based on simultaneous modeling of the process of losses and the process of data base construction. This approach is illustrated based on data related to operational risk losses.

Introduction

Consider a business (a bank, for example) that is interested in estimation of risks of a given type that it is facing. For example, banks are recently very interested in estimation of their exposure to so called *operational risk* that includes all risks except those related to markets and credit (some people consider this definition too general, but it is good enough for our purposes). We have at our disposal a database that contains description of operational losses suffered by various companies over a number of years; these losses are generally large and entries related to the bank of interest itself are extremely rare (otherwise the bank would not have stayed in business long enough to worry about risks!). This data base is in the initial phase of construction and is thus known to be incomplete: one can safely assume that it only refers to a small fraction of losses suffered by various businesses. The process of compiling the data base is typically focused on a certain *set of sources*; we will assume that only losses that appeared or might have appeared in this set of sources are relevant. Our main problem of interest is how to use such a database to gain information about the stream of losses facing the bank. One of our main points in addressing this problem is that for successful inference one must be able to model the process of compiling the data base - otherwise, as illustrated below, there is no objective way to characterize the process of losses. A summary of the Operational Risk issues can be found in the January 2002 issue of the *Risk* journal; in particular, see Alexander (2002).

1. The basic approach

We shall approach this problem in three stages. Our initial goal is to develop methods for characterizing the stream of losses related to operations observed worldwide and their magnitudes; this involves drawing inferences about the hidden population of losses that are not represented in

the data base. Subsequently, we will try to estimate what fraction of these losses is related to financial institutions; finally, we will estimate the parameters of the stream of operations related losses for a specific bank. The derived model is useful in several respects. First, it can be used by the bank to reserve capital needed to cover operational losses for a given period. Second, it can be used by insurance companies to assess the risk related to the bank and establish premiums. Though todays banks are mostly self-insured with respect to operational losses, there are reasons to expect that in the future many of them will prefer to handle this type of losses through insurance companies.

Estimation of properties of hidden populations has been considered in the literature in conjunction with such areas as Demography (e.g., population size estimation, see Rosenberg et al. (1995)), Software Reliability (estimation of the number of software defects hidden in the code, see Littlewood (1989)), or Non-destructive Evaluation (inference about hidden defects, see Meeker et al. (1996)). The corresponding techniques are referred to in the statistical literature as *size-biased sampling*. What makes the present problem special is its strong actuarial aspect: the questions that are asked in this context are very much different from those asked in the areas mentioned above, and these questions, in turn, determine the tools used in the statistical analysis. In essence, we are considering here a situation faced by every "young" branch of insurance when the data is sparse and expensive to collect, and the risks are poorly understood. It appears that the present day literature related to Actuarial Science does not provide an agreed upon statistical methodology for establishment of a new area. In this work, we will attempt to formulate a framework that could lead to such methodology.

Our basic assumptions are as follows:

- The process of losses is homogenous Poisson with rate λ events per year
- The underlying distribution of loss magnitudes is described by some density $f(x)$ that belongs to one of the families that are typically used to describe distribution of losses. For examples, Pareto, Weibull or Lognormal families can be considered good candidates.
- If a loss of magnitude x occurs, its probability of being discovered in the process of compiling a data base is $p(x)$, where p is a monotone function increasing from 0 to 1 with x. In essence, we demand that $p(x)$ satisfy the properties of a cumulative distribution function (cdf); in fact, we will use some of the cdf's stemming from applications that model growth to play the role of $p(x)$. In what follows we will refer to this function as the *Discovery Probability Curve (DPC).*

We note that in more complex applications the rate λ and parameters of the distribution of losses and the DPC will depend on a set of factors, as will be discussed in Section 6.

Consider the basic set of data shown in Appendix A. This data corresponds to losses compiled from public sources (news reports). A total of 226 cases were collected. For every case, the data gives the amount of loss in Deutsche Marks (DM) and the degree of relevance to the Banking Industry.

In our initial analysis we subdivided the data randomly into two parts, the learning sample and the test sample, as shown in the table. All the methods discussed below were first applied to the learning sample, and then validated on the test sample. In this article, however, we only show results for the overall sample. In Figure 1 we show the observed losses on the Weibull probability plot. The Weibull cdf $F(x)$ and density $f(x)$ are given by the extreme value distribution,

$$F(x) = 1 - exp\{-(x/b)^c\}, \qquad f(x) = (c/b)(x/b)^{c-1} exp\{-(x/b)^c\}, \quad x > 0 \tag{1.1}$$

and the estimated parameters of the law are $\hat{c} = 0.32$ and $\hat{b} = 4.9 \times 10^7$. From Figure 1 it appears as if the distribution is consistent with the Weibull ($\hat{c} = 0.32, \hat{b} = 4.9 \times 10^7$) law except that the smaller losses are missing, presumably because it is difficult for such losses to get into the set of sources and to be discovered.

On a closer look, however, such a simplistic explanation is hardly satisfactory. Suppose that the population of losses is indeed distributed in accordance with the above Weibull law. Then the fraction of operational losses below DM 1,000,000 in the overall population is estimated to be only 0.25, which does not make sense.

However, the fact that the Weibull probability plot is linear in the upper tail suggests that the upper tail of the distribution may indeed be Weibull (albeit with different parameters) and, with a suitably chosen and practically plausible DPC $p(x)$ we might get results that are consistent with the data in Appendix A.

To illustrate this point let us switch to logarithmically transformed data; in what follows we will work with the observations $y = ln(x)$. If the losses x_i are distributed in accordance with (1.1) then the cdf and density of log-losses are

$$\tilde{F}(y) = 1 - exp\{-exp[(y - u_1)/u_2]\}, \qquad \tilde{f}(y) = u_2^{-1} exp\{-exp[(y - u_1)/u_2] + (y - u_1)/u_2\} \quad (1.2)$$

(here and in what follows the "tilde" symbol will refer to quantities associated with log-losses). It is easy to see that

$$u_1 = ln(b), \qquad u_2 = 1/c. \quad (1.3)$$

Now let us define the DPC in terms of a logistic curve:

$$\tilde{p}(y) = \{1 + exp[-(y - v_1)/v_2]\}^{-1} \quad (1.4)$$

Let us now select the parameters $(u_1, u_2) = (10.2\ 7.5)$ and $(v_1, v_2) = (14\ 1.7)$. After simulating the process of losses and discovery in accordance with these parameters, we obtain a plot shown in Fig. 2. One could see that the observable data is very similar to that presented in Fig. 1. One can show, based on the methods presented later, that the particular distributions selected above do not contradict the data and thus can be considered (with some stretch, as we will see in Section 5) plausible. In the following section we develop methods for fitting models of this type.

2. The estimation problem

Let, in general, $\tilde{f}(y|\boldsymbol{u})$ and $\tilde{p}(y|\boldsymbol{v})$ are the density of log-losses and the DPC, respectively ($\boldsymbol{u}$ and $\boldsymbol{v}$ are the corresponding parameters). The density of a log-loss y conditional on this loss appearing in the body of sources and being discovered there is given by

$$\tilde{f}_c(y|\boldsymbol{u}, \boldsymbol{v}) = \{\tilde{f}(y|\boldsymbol{u})\tilde{p}(y|\boldsymbol{v})\}/C(\boldsymbol{u}, \boldsymbol{v}), \quad (2.1)$$

where the mean value of the discovery probability, represented by the normalizing constant $C(\boldsymbol{u}, \boldsymbol{v})$, is given by

$$C(\boldsymbol{u}, \boldsymbol{v}) = \int_{-\infty}^{\infty} \tilde{f}(y|\boldsymbol{u})\tilde{p}(y|\boldsymbol{v})dy. \quad (2.2)$$

Suppose that the overall number of losses recorded in the set of sources is N and the actual number of discovered losses is k; the corresponding log-losses are $y_1, y_2, \ldots, y_k$. Our main problem is estimation of the parameters $\boldsymbol{u}, \boldsymbol{v}$ and N.

2.1 Likelihood based estimation

The log-likelihood of the observed data is given by

$$L(\boldsymbol{u},\boldsymbol{v},N|y_1,y_2,\ldots,y_k) = ln\binom{N}{k} + k\times lnC(\boldsymbol{u},\boldsymbol{v}) + (N-k)\times ln(1-C(\boldsymbol{u},\boldsymbol{v})) + \sum_{i=1}^{k} ln\left[\tilde{f}_c(y_i|\boldsymbol{u},\boldsymbol{v})\right]; \tag{2.3}$$

the inference can now be based on this log-likelihood. When nothing else is known about the parameters, one can derive the Maximum Likelihood Estimators (MLE's) by finding the parameters that maximize (2.3). In this article we will not perform such a likelihood analysis; instead, we will work with a somewhat simplified form of the likelihood that arises when it is known a-priori that N is large and C is small. The presented approach adequately represents the main ideas and is sufficient to address the problems that motivated this research. Analysis of the exact likelihood (2.3) can be carried out in a similar way.

When it is known a-priori that N is large and only a small fraction of the losses is discovered, one can approximate the binomial term in (2.3) by the corresponding Poisson term. The approximate log-likelihood becomes

$$L(\boldsymbol{u},\boldsymbol{v},N|y_1,y_2,\ldots,y_k) \approx ln\left[\frac{(NC(\boldsymbol{u},\boldsymbol{v}))^k e^{-N}C(\boldsymbol{u},\boldsymbol{v})}{k!}\right] + \sum_{i=1}^{k} ln\left[\tilde{f}_c(y_i|\boldsymbol{u},\boldsymbol{v})\right], \tag{2.4}$$

In the process of Maximum Likelihood (ML) estimation one can take advantage of the fact that, for given $\boldsymbol{u}$ and $\boldsymbol{v}$ the likelihood is maximized when

$$N = k/C(\boldsymbol{u},\boldsymbol{v}), \tag{2.5}$$

indicating that one can expect to obtain estimates of good quality based on the conditional distribution of the observed losses only. After substitution of (2.5) into (2.4), we can obtain the estimates by solving the gradient equations,

$$\begin{cases} \sum_{i=1}^{k} \frac{\nabla_{\boldsymbol{u}}\tilde{f}(y_i|\boldsymbol{u})}{\tilde{f}(y_i|\boldsymbol{u})} = k \times \frac{\nabla_{\boldsymbol{u}}C(\boldsymbol{u},\boldsymbol{v})}{C(\boldsymbol{u},\boldsymbol{v})} \\ \sum_{i=1}^{k} \frac{\nabla_{\boldsymbol{v}}\tilde{p}(y_i|\boldsymbol{v})}{\tilde{p}(y_i|\boldsymbol{v})} = k \times \frac{\nabla_{\boldsymbol{v}}C(\boldsymbol{u},\boldsymbol{v})}{C(\boldsymbol{u},\boldsymbol{v})} \end{cases} \tag{2.6}$$

The equations (2.6) can be solved by using a simple iterative scheme that starts from some initial values $(\boldsymbol{u}_0,\boldsymbol{v}_0)$ and proceeds as follows:

Estimation Procedure A:

Step1: For the current values $(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{u}^{(i)},\boldsymbol{v}^{(i)})$ compute $C(\boldsymbol{u},\boldsymbol{v})$ and its gradient vectors by $\boldsymbol{u}$ and $\boldsymbol{v}$, $\nabla_{\boldsymbol{u}}C(\boldsymbol{u},\boldsymbol{v})$ and $\nabla_{\boldsymbol{v}}C(\boldsymbol{u},\boldsymbol{v})$.

Step2: Substitute the resulting values in the RHS of (2.6) and solve the two groups of equations separately. Assign the solutions to $(\boldsymbol{u}^{(i+1)},\boldsymbol{v}^{(i+1)})$.

Step3: Iterate Step 1 and Step 2 until the convergence occurs; Accept the result if it passes tests for local optimality, sanity and goodness of fit, as described later.

The tests for "sanity" mentioned in the procedure are needed because the solution of (2.6) maximizes the approximate log-likelihood (2.4) and not the exact likelihood, (2.3). Therefore, if for the resulting estimates $(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})$ the value $C(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})$ is not small enough to justify the Poisson approximation used to obtain (2.6), this solution cannot be considered acceptable. In such situations one cannot

expect that a small value of $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ will be obtained by solving the exact profile likelihood equations; therefore, failure of the equations (2.6) to produce a value of discovery probability, $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$, that is small enough to be compatible with one's expectation indicates that the full optimization approach is inadequate and some additional restrictions on parameters are necessary.

Once the estimates $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ are at hand, the estimate $\hat{N}$ is obtained by substituting these values into (2.5).

2.2 Constrained estimation and Inference

Many problems related to the model described above involve maximization of the likelihood in the presence of some constraints on the parameters. For example, after obtaining the ML estimates, one could decide that the resulting value of $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ is too high to be plausible, and carry out estimation under the constraint

$$C(\boldsymbol{u}, \boldsymbol{v}) = c_0, \tag{2.7}$$

where c_0 is a suitably chosen constant. The estimation can be carried out by introducing a Lagrange multiplier β associated with this constraint and maximizing the Lagrangian

$$L_\beta(\boldsymbol{u}, \boldsymbol{v}, N|y_1, y_2, \ldots, y_k) = L(\boldsymbol{u}, \boldsymbol{v}, N|y_1, y_2, \ldots, y_k) - \beta(lnC(\boldsymbol{u}, \boldsymbol{v}) - lnc_0). \tag{2.8}$$

One way to achieve this goal is to solve the gradient equations

$$\begin{cases} \sum_{i=1}^{k} \frac{\nabla_{\boldsymbol{u}} \tilde{f}(y_i|\boldsymbol{u})}{\tilde{f}(y_i|\boldsymbol{u})} = (k+\beta) \times \frac{\nabla_{\boldsymbol{u}} C(\boldsymbol{u},\boldsymbol{v})}{C(\boldsymbol{u},\boldsymbol{v})} \\ \sum_{i=1}^{k} \frac{\nabla_{\boldsymbol{v}} \tilde{p}(y_i|\boldsymbol{v})}{\tilde{p}(y_i|\boldsymbol{v})} = (k+\beta) \times \frac{\nabla_{\boldsymbol{v}} C(\boldsymbol{u},\boldsymbol{v})}{C(\boldsymbol{u},\boldsymbol{v})} \\ C(\boldsymbol{u}, \boldsymbol{v}) = c_0. \end{cases} \tag{2.9}$$

The above equations can be solved by repeating, for various values of β, the process similar to Procedure A until a value of β is found for which the constraint (2.8) is satisfied; the details of this algorithm will be omitted.

Another situation in which constrained optimization is used in conjunction with the likelihood analysis is when one is willing to assume that some components of the parameters are known. This will lead to a reduced system (2.6) which contains only the equations corresponding to unknown parameters; this system can be solved by using an approach described in Procedure A. For example, under the assumption that the vector $\boldsymbol{v}$ that characterizes the PDC is known and equal to $\boldsymbol{v}_0$, the estimation process boils down to solving the system

$$\sum_{i=1}^{k} \frac{\nabla_{\boldsymbol{u}} \tilde{f}(y_i|\boldsymbol{u})}{\tilde{f}(y_i|\boldsymbol{u})} = k \times \frac{\nabla_{\boldsymbol{u}} C(\boldsymbol{u}, \boldsymbol{v}_0)}{C(\boldsymbol{u}, \boldsymbol{v}_0)} \tag{2.10}$$

Constrained estimation also plays an important role in inference related to the parameters of interest. For example, let us assume that $\boldsymbol{v}$ is known and equal to $\boldsymbol{v}_0$, and one is interested in testing the hypothesis that $C(\boldsymbol{u}, \boldsymbol{v}_0) = c$ against the alternative $C(\boldsymbol{u}, \boldsymbol{v}_0) < c$, at the significance level γ. To achieve this goal, we can compute the maximum value of the log-likelihood under the constraint $C(\boldsymbol{u}, \boldsymbol{v}_0) = c$ (denote the constrained estimate by $\hat{\boldsymbol{u}}_c$) and reject the hypothesis if $C(\hat{\boldsymbol{u}}, \boldsymbol{v}_0) < c$ and

$$\Psi_C(c) = 2\left\{L\left[\hat{\boldsymbol{u}}, \boldsymbol{v}_0, k/C(\hat{\boldsymbol{u}}, \boldsymbol{v}_0)\right] - L\left[\hat{\boldsymbol{u}}_c, \boldsymbol{v}_0, k/C(\hat{\boldsymbol{u}}_c, \boldsymbol{v}_0)\right]\right\} > \chi^2_{1-\alpha}(1). \tag{2.11}$$

Furthermore, confidence bounds are obtained simply by collecting values that are not rejected by the corresponding test. For example, the value of c in the domain $c > C(\hat{\boldsymbol{u}}, \boldsymbol{v}_0)$ for which an equality is achieved in (2.11) represents a $(1-\gamma) \times 100\%$ upper confidence bound for $C(\boldsymbol{u}, \boldsymbol{v}_0)$.

As usual, two-sided $(1-\gamma)\times 100\%$ confidence interval is obtained by combining lower and upper $(1-\gamma/2)\times 100\%$ confidence bounds.

Likelihood based inference about $C(\boldsymbol{u},\boldsymbol{v})$ does not lead directly to inference about N. In particular, if $(\underline{C},\bar{C})$ is the $(1-\gamma)\times 100\%$ confidence interval for C then $(k/\bar{C}, k/\underline{C})$ does not provide enough coverage to serve as $(1-\gamma)\times 100\%$ confidence interval for N; however, these bounds are useful as initial points in the numeric procedure described below. To test the hypothesis that $N=n$ against the alternative $N<n$, we have to compute the maximum value of the log-likelihood under the constraint $N=n$. As can be seen from (2.4), this goal can be achieved by solving the gradient equations

$$\begin{cases} \sum_{i=1}^{k} \frac{\nabla_{\boldsymbol{u}} \tilde{f}(y_i|\boldsymbol{u})}{\tilde{f}(y_i|\boldsymbol{u})} = n \times \nabla_{\boldsymbol{u}} C(\boldsymbol{u},\boldsymbol{v}) \\ \sum_{i=1}^{k} \frac{\nabla_{\boldsymbol{v}} \tilde{p}(y_i|\boldsymbol{v})}{\tilde{p}(y_i|\boldsymbol{v})} = n \times \nabla_{\boldsymbol{v}} C(\boldsymbol{u},\boldsymbol{v}) \end{cases} \tag{2.12}$$

by using a suitably modified Procedure A. Denote the constrained estimates by $(\hat{\boldsymbol{u}}_n, \hat{\boldsymbol{v}}_n)$ and the score associated with N by

$$\Psi_N(n) = 2\left\{ L(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}},\hat{N}) - L(\hat{\boldsymbol{u}}_n,\hat{\boldsymbol{v}}_n,n) \right\}. \tag{2.13}$$

Then the hypothesis is rejected if $\hat{N}>n$ and $\Psi_N(n)>\chi^2_{1-\alpha}(1)$. The lower $(1-\gamma)\times 100\%$ confidence bound for N is then the value of $n<\hat{N}$ for which $\Psi_N(n)=\chi^2_{1-\alpha}(1)$. The upper bound is obtained in a similar way.

3. Goodness of Fit Tests

The fact that we successfully obtained estimates of the basic parameters does not mean much unless the data is compatible with our model. In this section we discuss methods that enable one to make a judgment about such compatibility. We will consider two situations. In the first one we assume that the population of losses, whether it fits the model or not, remains homogenous - in other words, we cannot readily point out sub-populations (SP) for which the underlying model parameters can be suspected to be different. In the second situation we have reasons to suspect non-homogeneity and will need to test whether this is indeed the case.

3.1 Homogenous population

When the only way in which the model does not fit is associated with the choice of a wrong model rather than with the presence of sub-populations, one can use a number of graphical and analytical tools to test the model adequacy. One important graphical tool is the probability plot. Denote the ordered observations (log-losses) by $y_{(1)}, y_{(2)}, \ldots, y_{(k)}$. Denote the Cumulative Distribution Function (cdf) of the observations, conditional on discovery, by

$$\tilde{F}_c(y|\boldsymbol{u},\boldsymbol{v}) = \int_{-\infty}^{y} \tilde{f}_c(t|\boldsymbol{u},\boldsymbol{v})dt = \left\{ \int_{-\infty}^{y} \tilde{f}(t|\boldsymbol{u})\tilde{p}(t|\boldsymbol{v})dt \right\} / C(\boldsymbol{u},\boldsymbol{v}) \tag{3.1}$$

Suppose that the estimates of the parameters are $(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})$. One form of a probability plot is obtained by plotting, for $i=1,2,\ldots,k$, the points $[i/(k+1), \tilde{F}_c(y_{(i)}|\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})]$. Failure of these points to form a straight line with slope 1 is an indication of lack of fit. Some standard tests, such as Kolmogorov - Smirnov test, can be used to test for significance of the observed lack of fit. It is recommended to use adjustments to account for the fact that the parameters have been estimated from the data, eg. see D'Agostino and Stephens (1986). Another form of the probability plot is sometimes useful in models involving special parametric structure, such as location-scale

equivariance. This form is obtained by computing the scores $s_i = \tilde{F}_c^{-1}[i/(k+1)]$ and plotting the points $(y_{(i)}, s_i), i = 1, 2, \ldots, k$.

Another useful method is to compare the log-likelihoods corresponding to individual losses against the expected values. Denote the mean and variance of a single log-likelihood term by

$$\begin{aligned} E(\boldsymbol{u},\boldsymbol{v}) &= \left\{\int_{-\infty}^{\infty} \tilde{f}(t|\boldsymbol{u})\tilde{p}(t|\boldsymbol{v}) \times ln\left[\tilde{f}(t|\boldsymbol{u})\tilde{p}(t|\boldsymbol{v})\right] dt\right\}/C(\boldsymbol{u},\boldsymbol{v}) - lnC(\boldsymbol{u},\boldsymbol{v}), \\ V(\boldsymbol{u},\boldsymbol{v}) &= \left\{\int_{-\infty}^{\infty} \tilde{f}(t|\boldsymbol{u})\tilde{p}(t|\boldsymbol{v}) \times ln^2\left[\tilde{f}(t|\boldsymbol{u})\tilde{p}(t|\boldsymbol{v})\right] dt\right\}/C(\boldsymbol{u},\boldsymbol{v}) - (E(\boldsymbol{u},\boldsymbol{v}) + lnC(\boldsymbol{u},\boldsymbol{v}))^2. \end{aligned} \tag{3.2}$$

Now assume that the parameters are equal to their estimated values. Then, under the assumption that the model is correct,

$$Z = \frac{\left\{\sum_{i=1}^{k} ln\left[\tilde{f}_c(y_i|\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})\right]\right\} - kE(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})}{\sqrt{k}V(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})}, \tag{3.3}$$

can be treated as a realization of a standard normal random variable. Therefore, we can reject, at the level of significance γ, the hypothesis that the observed losses come from the postulated model if $|Z| > z_{1-\gamma/2}$.

One can find a number of additional goodness-of-fit tests in D'Agostino and Stephens (1986).

3.2 Non-homogenous population

Consider, for example, the case where the data set contains losses corresponding to two types of businesses: Banking and Others. If we disregard distinction between the sub-populations and apply one of the tests described above, we might reach a conclusion that some given model adequately represents the observed losses. Yet, fitting two separate models to sub-populations of interest can explain the data much better. Suppose, for example, that we have identified m sub-populations $P_1, P_2, \ldots, P_m$ for which, as we suspect, the parameters of the underlying population of losses are different, but the DPC's are the same and are assumed to be known. We can then perform the test for homogeneity based on the following statistic:

$$T = 2\left\{\sum_{j=1}^{m} L(\hat{\boldsymbol{u}}_j, \boldsymbol{v}, \hat{N}_j|\boldsymbol{y}_j) - L(\hat{\boldsymbol{u}}, \boldsymbol{v}, \hat{N}|\boldsymbol{y})\right\}, \tag{3.4}$$

where

$\boldsymbol{y}_j$ is the sub-sample of losses corresponding to the j -th sub-population.

$\hat{\boldsymbol{u}}_j$ is the vector of estimated parameters based on the data for the j -th sub-population only.

$\hat{N}_j$ is the estimated number of losses in the j -th sub-population (the estimation is based on $\boldsymbol{y}_j$ only).

$L(\hat{\boldsymbol{u}}_j, \boldsymbol{v}, \hat{N}_j|\boldsymbol{y}_j)$ is maximum log-likelihood based on the data corresponding to the j -th sub-population only.

$\boldsymbol{y}$ is the overall sample

$\hat{\boldsymbol{u}}$ is the vector of estimated parameters based on the complete sample

$\hat{N}$ is the estimated overall number of losses

$L(\hat{\boldsymbol{u}}, \boldsymbol{v}, \hat{N}|\boldsymbol{y})$ is maximum log-likelihood based on the complete sample

If the population is homogenous, the statistic T should have a chi-square distribution with the number of degrees of freedom equal to the product of m and the number of parameters in which the sub-populations differ from each other; for example, if the distributions of log-losses corresponding to different populations can differ in both location and scale, the number of degrees of freedom is $2m$. We reject the homogeneity hypothesis at the level of significance γ if T exceeds the $(1-\gamma)\times 100\%$ -th quantile of the chi-square distribution mentioned above.

It is not difficult to generalize the above test for the case where the PDC parameters for various sub-populations can also be different.

4. Tail Based Inference

One can still take advantage of the model under consideration even if the goodness of fit tests based on the complete data set suggest its rejection. Consider the situation where the company has to estimate reserves needed to cover the overall losses in the coming year. Consider two types of losses: small losses (not exceeding some prescribed level A) and large losses (greater than A). The company has enough internal information to estimate the magnitude and frequency of small losses. Larger losses, however, are observed rarely within the company, providing no solid basis for statistical estimation. It is then natural to perform that data analysis under the working assumption that the distribution of large losses pertaining to the company's business can be estimated based on observed losses suffered by "similar" companies. The company performs a search of the body of sources to collects information on such losses. Suppose that most of the discovered losses are greater than A and our attempt to fit a model involving Weibull losses and Logistic PDC fails; there is still a possibility that this model will fit to a suitably transformed data if we limit our attention to the population of losses that are greater than A. For example, such a model could fit well some form of an excess loss data, such as $(x_i - A)$ or $ln(x_i/A - 1)$.

Another area in which estimation in the domain $x > A$ is of primary interest is insurance. Suppose that the company intends to insure itself against losses exceeding A (here A could also represent the deductible demanded by the insurance company. From the insurance company point of view, losses below A are of no interest, and its risk analysis can be performed solely based on a distribution that fits the data only in the domain $x > A$.

Instead of fitting some distribution to some form of excess loss data as suggested above, one could use an alternative approach inspired by the asymptotic theory of sample extremes (see Galambos (1987)). One of the main subjects of this theory is analysis of distributions that have a Pareto tail index, i.e.,

$$1 - F(x) \sim x^{-a} L(x), \text{ as } x \to \infty, \tag{4.1}$$

where $L(x)$ is some *slowly varying* function, i.e., a function that satisfies the relation

$$L(tx)/L(x) \to 1 \text{ as } x \to \infty, \tag{4.2}$$

for every $t > 0$. This class is very extensive and it includes many of the distributions used by practitioners to model losses. When A is a large number (as in the case in insurance applications or the problem of operational risk estimation described in the beginning of the section), the distribution of the data in the domain $x > A$ is given by

$$F(x|x > A) = 1 - (x/A)^{-a}[L(x)/L(A)] \approx 1 - (x/A)^{-a}, \quad x > A. \tag{4.3}$$

The above approximation, suggested by (4.2), can be justified in many practical situations. In the simplest case where the distribution of losses in the range of interest $x > b$ is a two-parameter Pareto,

$$F(x) = 1 - (x/b)^{-a}, \quad x > b, \tag{4.4}$$

i.e., $L(x) \equiv 1$, the approximation in (4.3) reduces to equality. In terms of logarithms, the distribution becomes shifted exponential, i.e.,

$$\tilde{F}(y|y > u_1) = 1 - exp\{-(y - u_1)/u_2\}, \qquad \tilde{f}(y|y > u_1) = u_2^{-1} exp\{-(y - u_1)/u_2\}, \quad y > u_1 \tag{4.5}$$

where

$$u_1 = ln(A), \qquad u_2 = 1/a. \tag{4.6}$$

The above argument illustrates the point that in the tail area the location-scale distribution families once again lead to relatively tractable models; however, the location parameter typically turns out to be the left endpoint of the corresponding distribution.

In general, the problem of inference in the case where the log-losses are treated as left-censored, are assumed to come from the distribution $\tilde{f}(y|y > u_1; \boldsymbol{u})$, and are observed in accordance with some DPC $\tilde{p}(y_i|\boldsymbol{v})$, is similar to that described in Section 3. From the practical standpoint, this analysis is frequently simpler because u_1 can be treated as known.

5. Examples

Consider the data in Appendix A. To illustrate application of the described methods, considering two cases: in the first case we fit the Weibull-Logistic (WL) scheme to the whole distribution of losses contained in the set of sources. In the second case we focus on large losses (those exceeding some "deductible" or other boundary of interest) exclusively, disregarding possible lack of fit for the distribution as a whole.

5.1 Global Weibull-Logistic Model

In this section we assume that the underlying distribution of losses is Weibull, i.e., the log-losses are distributed in accordance with (1.2) and that the DPC is logistic (1.4). In the first phase, let us estimate the parameters $(\boldsymbol{u}, \boldsymbol{v}, N)$ without imposing any restrictions on them. Maximization of the log-likelihood (2.4) leads to the estimates $\hat{\boldsymbol{u}}$ = (17.1, 3.74) and $\hat{\boldsymbol{v}}$ = (8.78, 0.34), which imply, by (2.2), that $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}) = 0.90$ and, by (2.5), that $\hat{N} = 225/0.90 = 250$. In other words, the "best" explanation of the data offered by an unconstrained model is as follows: the losses are coming from the Weibull distribution with a very large scale parameter, $exp(17.1) = 2.67 \times 10^7$ and not very small shape parameter, $c = 1/3.74 = 0.27$. So, the underlying set of sources does not contain many small and moderate losses. The loss corresponding to the probability of discovery 0.5 is $exp(8.78) = 6500$ and losses corresponding to the probability of discovery 0.1, 0.25, 0.75, 0.9 and 0.99 are given in the first line of Table 1. Basically, the unconstrained approach suggests that all losses except those below 30000 are discovered and represented in the data at hand.

From the practical standpoint this explanation, of course, does not make any sense and the above example illustrates what could happen if one does not think a-priori about the plausible values of the parameters and counts on statistical estimation to "find" them. In the next step, assume that one has reasons to believe that the DPC parameters can be treated as known and equal to $\hat{\boldsymbol{v}}_0$ = (14 1.7). This suggests that the loss corresponding to the probability of discovery

0.5 is 1.2M and losses corresponding to other values of the probability of discovery are given in the second line of Table 1. Use of the constrained estimation procedure described in Sec. 2.2 leads to the estimate $\hat{\boldsymbol{u}}$ = (10.2, 7.5) which appears more sensible from the practical point of view: now the large losses are explained not as much by a *large scale* parameter, as in the unconstrained case, but by a *smaller shape* parameter. This suggests that the bulk of the losses in the body of sources still lay undiscovered: the estimated proportion of discovered losses is $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}_0) = 0.20$ and, consequently, the estimated number of losses recorded in the set of sources is $\hat{N} = 225/0.20 = 1125$.

An noted in the Introduction, a simulated sample from this model is shown in Figure 2. Even in the unconstrained case, there is no assurance that the model corresponding to estimated parameter values will fit the data (in fact, it does, as the interested reader can verify by applying the tests illustrated below). Once constraints are imposed, it becomes quite possible that the model will fit poorly and careful examination of goodness of fit issue is in order. We apply some of the techniques described in Section 3 to verify that the constraint $\boldsymbol{v}_0$ = (14 1.7) is compatible with the data. First, let us consider the probability plot (see Figure 3). The maximal deviation from the straight line, is 0.09. Practitioners frequently use 0.05 as the cut-off value, so we would reject the hypothesis that the model fits the data if this deviation exceeds the critical 5% value for the Kolmogorov-Smirnov statistic. Since this critical value is known to be $1.36/\sqrt{225} = 0.09$, we have no sufficient evidence that the model does not fit the data.

To apply a test based on the likelihood, note that the log-likelihood of the data in the constrained model is -605. Formulas (3.2) suggest that for data coming from the model with parameters $\hat{\boldsymbol{u}}$ = (10.2, 7.5) and $\hat{\boldsymbol{v}}$ = (14 1.7) the average score per observed loss is $E(\boldsymbol{u}, \boldsymbol{v})$ = -2.63 and the variance is $V(\boldsymbol{u}, \boldsymbol{v}) = 0.55$. This suggests that the value of the log-likelihood observed under the estimated model is approximately normal with mean and standard deviation -2.62×225 = -591 and $\sqrt{0.55 \times 225} = 11.1$, respectively. The value -605 is within 1.27 standard deviations from the mean, which corresponds to the p-value of 0.1 for the one-sided goodness of fit test - so, this test also does not lead to rejection of the model.

One could notice, however, that the constrained model barely squeaks by, and it does not even have some features that a practitioner may desire: given that the provided data is a result of a limited search effort, the probability 0.1 of discovering a loss of magnitude 28000 in the body of sources appears too high and the overall probability of discovery 0.2 appears way too high as well. However, our analysis shows that an attempt to obtain a much better Weibull-Logistic model compatible with the data set at hand does not lead anywhere: models that appear more attractive from the practical standpoint unfortunately do not fit, especially in the lower end. The difficulties are probably related to the fact that efforts of putting together the data base are in the very initial stage and appear very uneven in coverage. Furthermore, some values of the data have a much higher probability that the neighboring data, which exposes the fact that a Weibull model is a-priori just a convenient mathematical approximation: for example, a typical small fine imposed by a judge against an operational risk related violation and reported in the press is much more likely to be \$10,000 (DM 18497 in the data set) than \$9000. Though such partial grouping does not derail the estimation process, it is advisable to establish, in every individual study, what is its effect on both estimation and goodness of fit tests.

Finally, we test whether, under the assumption that the model is WL with $\hat{\boldsymbol{v}}$ = (14 1.7), the population of losses classified as being of "high" or "medium" relevance (sub-population 1, denoted SP1) differs significantly from the population classified as being of "low" relevance (sub-population 2). Let us fit two separate models for the two sub-samples. The estimated population parameters (based on sample size of 140) for SP1 are $\hat{\boldsymbol{u}}_1$ = (9.8, 7.9) and the maximal value of the log-likelihood is -385 (the model cannot be rejected by goodness of fit test, but the quality of fit is marginal). The parameters for SP2, based on sample size of 85, $\hat{\boldsymbol{u}}_1$ = (10.7, 7.0). The maximal value of the

log-likelihood is -220 and the fit is very good. As indicated earlier, the maximum log-likelihood value for the complete data set was -605. To test whether the complete data set is explained better by two separate models for SP1 and SP2 than by a single model, we need to compute T. Since we have two sub-populations that differ in two parameters, T should be compared to $\chi^2_{0.95}(4) = 9.49$. In our case $T = 2(-385 - 220 + 605) = 0$, indicating that we have no evidence, under the given PDC, to conclude that there is a significant difference between the loss distributions corresponding to SP1 and SP2.

5.2 Tail Pareto-Logistic (PL) Model

Now let us consider the situation from the prospective of the insurance company and assume that only losses exceeding A = DM 1.2M (the deductible) are of interest. Assume that the distribution of losses is two parameter Pareto; as noted in Section 4, this implies that the log-losses are distributed in accordance with (4.5) with $u_1 = ln(1.2 \times 10^6) = 14$. The only parameters of interest are $\boldsymbol{v}$, the scale u_2 and N. The relevant data is now reduced from 225 to 163 losses exceeding 1.2M.

First, let us apply the unconstrained model. Maximization of the likelihood based on *conditional* density functions leads to the estimates $\hat{\boldsymbol{u}}_2 = 1.14$, $\hat{\boldsymbol{v}} = (19.7, 0.98)$. Therefore, $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}) = 0.02$ and the total number of losses exceeding 1.2M contained in the set of sources is estimated to be $\hat{N} = 163/0.02 = 8150$. This model suggests that the loss that has the probability 0.5 to enter into the set of sources and be discovered is 3.7×10^8). Losses corresponding to various values of the DPC are shown in Table 1.

Once again, one has a reason to be disappointed in the results of the automatic search for the model parameters: it appears as if losses of very high magnitude have a high chance to be overlooked in the process of building the data base. One could "correct" this anomaly by introducing a constraint that the parameters of the Logistic DPC are $\boldsymbol{v}_0 = (17, 1)$. Under this constraint the magnitude of a loss that has a 0.5 probability to be discovered is 2.4×10^7); other values (see Table 1), also appear to be more reasonable to a decision maker. The resulting estimate of the scale is $u_2 = 1.97$. The estimated overall discovery probability is now much larger, namely, $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}) = 0.30$ leading to $163/0.3 = 543$ as the estimate of total number of losses of magnitude exceeding 1.2M in the time period of interest.

To check whether the resulting Pareto-Logistic model (for losses exceeding 1.2M) with parameters $\hat{\boldsymbol{u}} = (14, 1.97)$, $\hat{\boldsymbol{v}} = (17, 1)$ fits the data, we first examine the probability plot in Figure 3: the Kolmogorov-Smirnov test suggests that the fit is good. The mean and variance of the log-likelihood score corresponding to a single measurement are $E(\boldsymbol{u}, \boldsymbol{v})$ = -2.19 and $V(\boldsymbol{u}, \boldsymbol{v}) = 0.55$. Therefore, the mean and standard deviation of the log-likelihood under the assumption of the above model are -2.19×163 = -356 and $\sqrt{0.55 \times 163} = 9.4$. The maximal log-likelihood computed for the data at hand under the constraint $\boldsymbol{v} = \boldsymbol{v}_0 = (17, 1)$ is -353, which is in agreement with the mean and standard deviation computed above.

6. Discussion

The limited scope of the data set discussed in this article enables one to answer questions of type discussed above. However, one would need a much more elaborate data set in order to address more complex questions. For example, consider the problem of building a model to estimate operational risk losses for a given enterprise. A data base suitable for such estimation would consist of a list of losses; for each record we would have not only loss magnitude and relevance, but also such entries as Industry, Number of Employees, Type of Loss, Market Value of the bank, etc. A promising strategy is to:

- establish, for each factor, whether it is affecting (i) λ, (ii) (u_1, u_2) or (iii) (v_1, v_2); in some

cases the type of an effect can be reasonably well postulated - for example, for some types of losses the rate could be assumed to be roughly proportional to the number of employees. Such a-priori relationships can considerably simplify the subsequent analysis. Their validity can be tested by using post-estimation goodness-of-fit procedures.

- Estimate the relationship between factors and the basic model parameters, λ, (u_1, u_2), (v_1, v_2)
- For a given enterprise P evaluate, based on the above model, the corresponding parameters, λ_P, (u_{1P}, u_{2P}), (v_{1P}, v_{2P}).
- Evaluate risks related to P based on the estimated parameters

We intend to consider such more general models in future research.

Acknowledgements

I deeply appreciate help by Dr. Mark Laycock (Deutsche Bank) and Dr. Stephen Witter (Deutsche Bank) for providing comments and criticisms on the original version of the manuscript, and to Dr.s Jonathan Hosking, Dirk Siegel and Katie Richards (IBM) for useful discussions on this subject.

REFERENCES

1. Alexander, C. (2002). Rules and Models, *Risk,* , No. 1, pp. 18-20.
2. D'Agostino, R.B. and Stephens, M.A. (1986). *Goodness-of-Fit Techniques*, Marcel Dekker, New York.
3. Galambos, J. (1987) *The Asymptotic Theory of Extreme Order Statistics*. Robert E. Krieger Publishing Co., Malabar, Florida.
4. Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, 2 vols., 2nd Ed., Wiley, New York.
5. Klugman, S.A., Panjer, H.H. and Willmot, G.E. (1998). *Loss Models*, Wiley, New York.
6. Littlewood, B. (1989). Predicting Software Reliability, *Phil. Trans. Royal Soc. London, A*, Vol. 327, pp. 513-527.
7. Olin, B.D. and Meeker, W.Q. (1996). Applications of Statistical Methods to Nondestructive Evaluation, *Technometrics*, Vol. 38, No. 2, pp. 95-112.
8. Rosenberg, D.K. and Overton, W.S. (1995). Estimation of Animal Abundance when Capture Probabilities are Low and Heterogenous. *J. Wildlife Management*, Vol. 59, pp. 252-261.

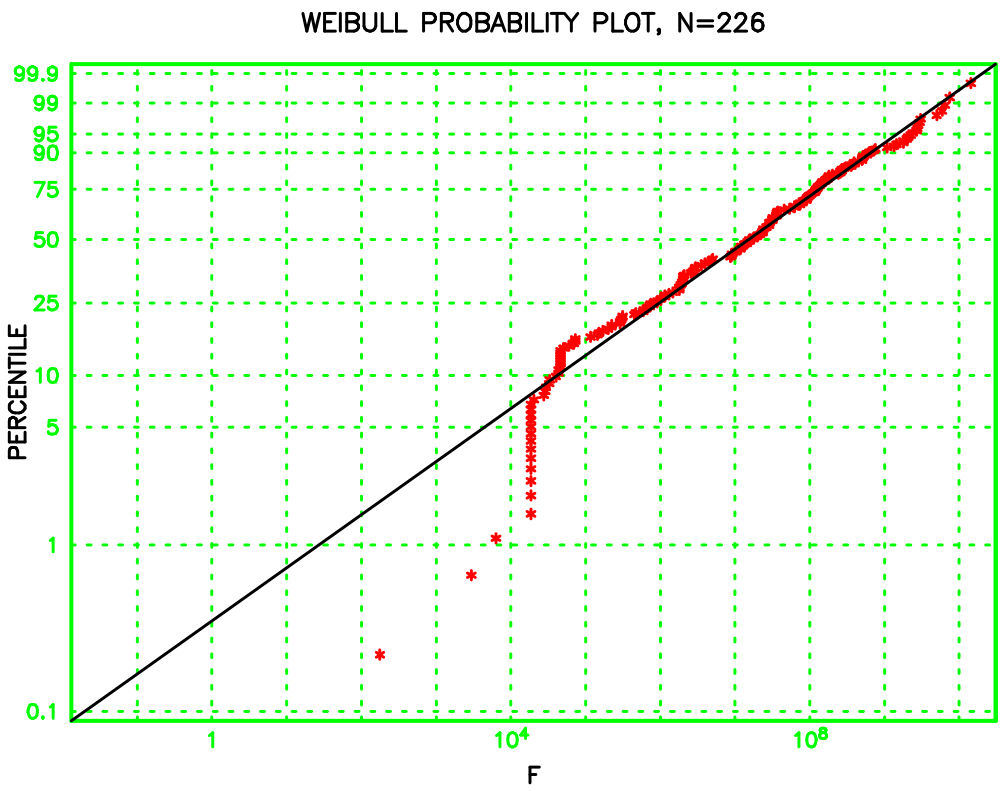

Figure 1. Observed losses plotted on the Weibull probability plot.

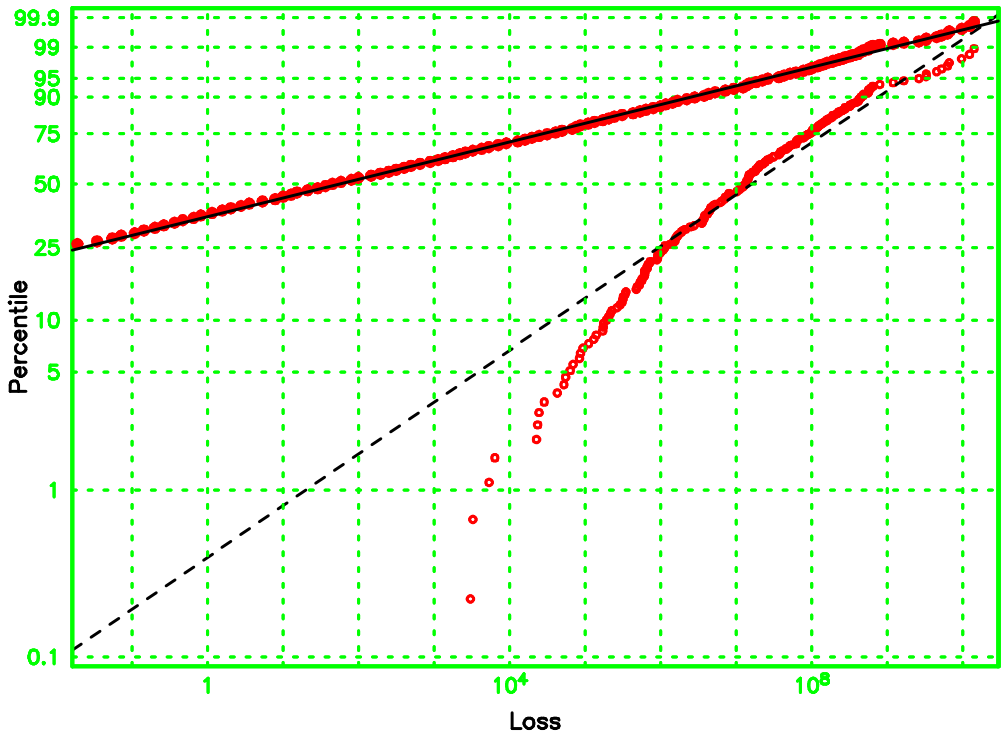

Figure 2. Simulated replica of the data obtained by combining Weibull distribution of losses and a logistic PDC.

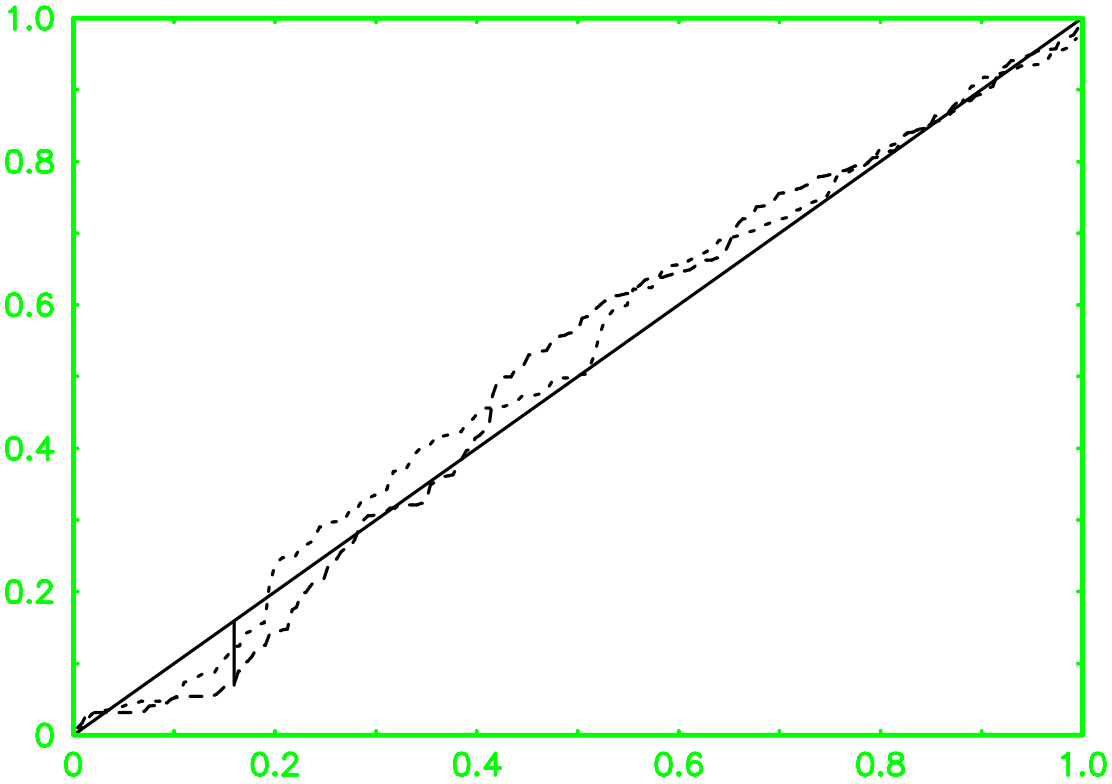

Figure 3. Probability plots: Weibull-Logistic (WL) model, with parameters $\boldsymbol{u}$ = (10.2, 7.5), $\boldsymbol{v}$ = (14, 1.7), dashed line, and Pareto-Logistic (PL) model with parameters $\boldsymbol{u}$ = (14, 1.98), $\boldsymbol{v}$ = (17, 1) in the domain $y > 14$, dotted line. The maximal Kolmogorov-Smirnov deviation is shown for the WL model.

	0.1	0.25	0.5	0.75	0.9	0.99
WL, $\boldsymbol{u}$ = (17.1, 3.74), $\boldsymbol{v}$ = (8.78, 0.34):	3100	4500	6500	9400	14000	31000
WL, $\boldsymbol{u}$ = (10.2, 7.5), $\boldsymbol{v}$ = (14, 1.7):	28000	1.9×10^5	1.2×10^6	1.8×10^6	5.0×10^7	3.0×10^9
PL, $\boldsymbol{u}$ = (14, 1.14), $\boldsymbol{v}$ = (19.7, 0.98):	4.2×10^7	1.2×10^8	3.7×10^8	1.1×10^9	3.2×10^9	3.4×10^{10}
PL, $\boldsymbol{u}$ = (14, 1.98), $\boldsymbol{v}$ = (17, 1):	2.7×10^6	8.1×10^6	2.4×10^7	7.2×10^7	2.2×10^8	2.4×10^9

Table 1. Values of DPC for Weibull-Logistic (WL) and Pareto-Logistic (PL) models.

APPENDIX A: Data used in the Example

.

Company	Relevance	Loss
ABN-Amro	High	126000000
AIG	High	162000000
Airtours	Low	23200000
Albatross-Warehousing	High	600000
Allied-Colloids	Low	32726.5
Allied-Colloids	Low	50300.5
Allied-Colloids	Low	469800
Allied-Colloids	Low	59458.7
Allied-lyons	High	450000000
Anheuser-Busch	Medium	63000000
ARCO-Pension-Fund	High	39600000
Asesores-de-Valores-(AVA)	Medium	177
AsiaFocus-and-others	Medium	12132000
Askin-Securities	High	1080000000
Astra-USA	High	17730000
B-Pacoroni	High	29500
Baii-Asset-Management	High	300000
Banco-Bilbao-Vizcaya	High	1000235.349
Bank-Of-America	High	75000000
Bankers-Trust	High	8694000
Bankers-Trust	High	240000000
Banque-Nationale-de-Paris	High	31500000
Barings	High	2400000000
Barloecher	Low	117000000
Baxter-Healthcare	Low	28600000
Bear-Stearns	High	46242.5
Bedfordshire-County-Council	Medium	4620000
Bent-Emanuel-Christiansen-of-Saeby	Low	737500
Bingdon-Builders	Low	29000
Boliden	Low	10400000
BP	Low	28620000
Bre-X	Low	7500000000
Britanic-Assurance	High	1950000
British-Airways	Low	362500000
British-Airways	Low	9025000
British-Airways	Low	912500
British-Government	Low	6380000000
British-Petroleum	Low	290000000
British-Rail	Low	72500
Bula-Resources	Low	36900000
Butte-Mining	Low	75000000
Campbell-Soups	Low	1160000
Canadian-Government	Low	20800000
Cantrade	High	118000000
Capel-Cure-Myers	High	24000000
Cargill-(Minnetonka-Fund)	High	180000000
Caterpillar-Financial	High	14940000
Centrica	Low	2250000000
Chiroscience	High	10075000

City-Technology-Holdings	Low	14500000
Codelco	High	360000000
Commercial-International-Bank	High	9860000
Connex-South-Eastern	Low	5015000
Connex-South-Eastern	Low	3944000
Credit-Commercial-de-France	High	23859230.54
Credit-Suisse	High	224000
Credit-Suisse	High	900000
Credit-Suisse-First-Boston	High	95940000
Credit-Suisse-First-Boston	High	1782000
Credit-Suisse-First-Boston	High	1782000
CreditLyonnais	High	31500000
Cruden-Construction	Low	310300
Cynamaur	Low	150800
Daewoo	Medium	11570000
Daimler-Benz	Low	600000000
Dain-Bosworth	High	2034000
Daiwa-Bank	High	1980000000
Dean-Witter-Reynolds	High	40000
Dell-Computer	High	62280000
Detroit-Edison	Medium	74250000
Deutsche-Bank	High	200000
Diocese-of-Dallas	Low	9000000
Dolophin-Drilling	Low	2950
Dresdner	High	70000
Dura-Automotive-Systems-Inc	Low	11700000
Dura-Automotive-Systems-Inc	Low	1800000
Endessa	High	3530242.41
Equitable-Life	High	270000000
Eurotunnel	*	580000000
Exxon	Low	14400000000
First-of-America-Securities	High	18497
First-Southwest-Company	High	18497
Florida-State-Treasury	High	1792800000
Garibaldi-Small-goods-Py-limited	Low	13230000
General-Accident	High	1101000
Gibson-Greeting	High	35460000
Glaxo	High	393000000
Go-Ahead-Group	Low	2082200
Goldman-Sachs	High	58800000
Goldman-Sachs	High	46242.5
Goldman-Sachs	High	138727500
Goldman-Sachs	High	101700000
Goldman-Sachs	High	6300
Goldman-Sachs	High	2034000
Grant-Thornton	High	43500
Great-Eastern-Railway	Low	2593050
Hammersmith-and-Fulham-Council	High	1800000000
Harris-Trust-and-Savings-Bank	High	92340000
Hickson-and-Welch	Low	1160000
Hickson-and-Welch	Low	145000
Hoover	Low	144000000
Hyundai	Low	29510000

Imperial-College-of-Science-Technol.	Low	450000
Intel	Low	1440000000
Invesco	High	33000000
JH-Marsh-McLennan	High	590000
J-P-Morgan-Securities	High	46242.5
John-Sisk-Sons	Low	3900000
Jones-Day-Reavis-Pogue	High	91800000
JP-Morgan	High	3292500
Kasima-Oil	High	2700000000
Kaye-ScholerFierman-Hays-Handler	High	73800000
Kensey-Nash	Low	170280
Kidder-Peabody	High	630000000
Kidder-Peabody	High	18000000
KN-Kwikform	Low	116000
KPMG-Peat-Marwick	High	135000000
Kraft-Foods-Limited	Low	12000000
LW-Insulations-Ltd	Low	20300
Lazard-Freres	High	2034000
LeBoeuf-Lamb-Greene-MacRae	High	104760000
Lehman-Bros	High	2034000
Lehman-Brothers	High	147976000
Lehman-Brothers-Imternational	High	120000000
Lenzing	Low	19675000
Lenzing	Low	2175000
LG	Low	13260000
Liberty	Medium	716300
Lloyds-TSB	High	2079750000
London-and-Manchester-Assurance	High	1950000
LucasVarity	Low	31900000
Mattel-Inc	Low	36000000
Mead-Corp	High	21780000
Medani	High	90000000
Merril-Lynch	High	180000000
Merril-Lynch	High	175721500
Merril-Lynch	High	290000
Merrill-Lynch	High	91000000
Merrill-Lynch	High	54000000
Merrill-Lynch	High	720000000
Merrill-Lynch-Pierce-Fenner-Smith	High	18497
Mersey-Docks-and-Harbour-Co	Low	30000000
Metallgesellschaft	High	2762000000
MGM/UA-Communications-CO	Low	64739500
Microsoft	Low	15487500
Miller,-Johnson-Keuhn	High	18497
Morgan,-Keegan-Co	High	18497
Morgan-Stanley-Co	High	18497
National-Express	Low	1475000
Nations-Bank	High	12150000
Natl-Assoc-of-Securities-Dealers	High	160000000
NatWest	High	270000000
Nelson-Wheeler	Medium	127800000
Neville-Russell	High	43500
Nortehrn-Railroad-Company	Low	2774550

Occidental-Petroleum	Low	5900000000
Oppenhiemer-Co	High	18497
Orange-County	High	3060000000
PO-Nedlloyd	Low	81075900
Pacific-Horizon-Funds	High	122220000
PaineWebber	High	46242.5
Pan-Am	Low	37044300
Paramount-Communications	High	36000000
Phillip-Morris	Low	98000000
Piper-Jaffray	High	18497
Piper-Jaffray	High	2034000
PNC-Capital-Mkts	High	18497
Powerscreen-International	Low	173365000
Prudential-Corporation	High	1350000000
Prudential-Insurance-of-America	Medium	1849700
Prudential-Insurance-of-America	Medium	2959520000
Prudential-Securities-Inc	High	46242.5
Quadrex-	High	29000000
Quilter-Fund-management	High	660000
Railtrack	Low	222430
Raymond-James-Assoc	High	18497
Rhone-Poulenc	Low	23490000
RMC-Group	Low	14413000
Royal-Bank-of-Scotland	High	255000000
Royal-Bank-of-Scotland	High	1299000
Salomon-Briothers	High	504000000
Salomons-Smith-Barneyn	High	129479000
Samsung	Low	14820000
Samuel-Montagu	High	498800000
Schwab	High	31444900
Seattle-Northwest-Securities	High	18497
Sedgwick	High	23200000
SG-Warburg	High	109800000
Shell	Low	258000000
Showa-Shell-Sekiyu	High	2844000000
SK-Group	Low	25350000
Smith-Barney	High	46242.5
Sony-Corporation	High	1681000
South-East-Infrastructure-Maint.-Co	Low	314360
South-West-Trains	Low	2938200
Southern-Track-Renewals	Low	27550
Soverign-Unit-Trust	High	4875000
Spicer-Oppenheim-(Delloitte-Touche)	Medium	290000
Stagecoach	Low	7375000
Stagecoach	Low	2882600
Stamford-Tyres	Low	36000000
Standard-Chartered-Bank	High	28000000
Stone-Youngberg	High	18497
Sumitomo	High	4994100000
Sumitomo-Finance-International	High	4500000
Sun-Life-of-Canada	High	21750000
SunTrust-Capital-Mkts	High	18497
Sutro-Co	High	46242.5

TACA	Low	506459100
TI-Group	Medium	26460000
Tilbury-Douglas	Low	72500
Tom's-of-Maine	Low	720000
Transpot-Life-Insurance	Low	45180000
Trevor-Osborne-Property-Group	Low	8990000
UEM	Medium	33000
Union-Bank-of-Switzerland	High	750000000
Union-Bank-of-Switzerland	High	384000000
UPS	High	145260000
Virgin	Low	1875000
Volkswagen	Low	201450000
Warwickshire-Health-Authority	Low	2900000
West-Bromwich-Building-Society	Medium	29000000
West-Indies-Cricket-Board	Low	2700000
Winchester-Commodities	High	301000000
Wood-Gundy	High	18000000
Yakult-Honsha	High	517600000