MINIMUM DISTANCE ESTIMATION OF LOSS DISTRIBUTIONS

STUART A. KLUGMAN AND A. RAHULJI PARSA

Abstract

Loss distributions have a number of uses in the pricing and reserving of casualty insurance. Many authors have recommended maximum likelihood for the estimation of the parameters. It has the advantages of asymptotic optimality (in the sense of mean square error) and applicability (the likelihood function can always be written). Also, it is possible to estimate the variance of the estimate, a useful tool in assessing the accuracy of any results. The only disadvantage of maximum likelihood is that the objective function does not relate to the actuarial problem being investigated. Minimum distance estimates can be tailored to reflect the goals of the analysis and, as such, should give more appropriate answers. The purpose of this paper is to demonstrate that these estimates share the second and third desirable qualities with maximum likelihood.

1. DEFINITIONS, NOTATION, AND AGENDA

We start with a definition of a minimum distance estimate. Let $G(c;\boldsymbol{\theta})$ be any function of c that is uniquely related to $f(c;\boldsymbol{\theta})$, the probability density function (pdf) of the population. By uniquely related we mean that if you know f, you can obtain G and vice versa. Call G the model functional. Let $f_n(c)$ be the empirical density. It assigns probability $1/n$ to each of the n observations in the sample. Let $G_n(c)$ be found from f_n in the same way that G is from f. Call G_n the empirical functional. The objective function is

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{k} w_i \left[G(c_i; \boldsymbol{\theta}) - G_n(c_i) \right]^2 , \tag{1.1}$$

where $c_1 < c_2 < ... < c_k$ are arbitrarily selected values and $w_1, w_2, ..., w_k > 0$ are arbitrarily selected weights. The weights can be selected either to minimize the variance of the estimate or to place emphasis on those values where a close fit is desired. The c_i will almost certainly be the class boundaries for whatever grouping was used in the initial presentation of the data. The minimum distance estimate is the value of $\boldsymbol{\theta}$ that minimizes $Q(\boldsymbol{\theta})$.

There are two functionals that appear to be appropriate for casualty work. The first is the limited expected value (LEV) which is useful in ratemaking. It is the expected loss when losses are capped at a specified value. This quantity is fundamental for calculating deductibles, limits, layers, increased limits, or the effects of inflation. This quantity is also useful for reserving if information about the distribution of outstanding claims is desired. Many practitioners make it a point to verify that the model LEVs (after estimating the parameters by maximum likelihood) and the empirical LEV match. Using the LEV as a distance measure gives this the best chance of happening.

The specific relationships are (when dealing with the LEV we will use L in place of G):

$$L(c; \boldsymbol{\theta}) = \int_0^c x f(x; \boldsymbol{\theta}) dx + c \int_c^\infty f(x; \boldsymbol{\theta}) dx \tag{1.2}$$

and

$$L_n(c) = {}^1\!/_n \sum_{i=1}^{n} \min (x_i , c) . \tag{1.3}$$

It should be noted that to compute $L_n(c_i)$ all that is needed is the number of observations, n_i, that are between c_{i-1} and c_i (where $c_0 = 0$) and the average, a_i, of these observations. Then

$$L_n(c_i) = \left[\sum_{j=1}^{i} n_j a_j + c_i \left(n - \sum_{j=1}^{i} n_j\right) \right] / n = c_i + \sum_{j=1}^{i} n_j (a_j - c_i)/n \,. \tag{1.4}$$

A second functional, one that makes sense for loss reserving, is the distribution function. As will be seen in the second example, loss distributions can be used to estimate the number of incurred but not reported (IBNR) claims. The key to the calculation is that the distribution function is evaluated at the highest lag for which losses have been reported. Using F for G we have

$$F(c; \boldsymbol{\theta}) = \int_0^c f(x; \boldsymbol{\theta}) dx \tag{1.5}$$

and

$$F_n(c) = \tfrac{1}{n} \, (\text{number of } x_i \le c) \,. \tag{1.6}$$

There are a number of steps that need to be taken to make this method practical.

1. Techniques for minimizing Q.
2. Verification that the solution possesses desirable statistical properties. This would include being unbiased, consistent, and, if not minimum variance, at least providing for calculation of the variance.
3. A demonstration that estimators obtained from this method are not unlike those obtained by maximum likelihood, at least when the data actually come from the distribution family being fitted.
4. Construction of a hypothesis test based on Q. This would allow for verification that the model selected is reasonable as well as for comparison with competing models.

This paper addresses Issues 1 and 2 in full and makes a proposal relative to Issue 4. The third issue requires a fairly substantial simulation, something we have elected not to do at this time. This paper includes two examples and a small simulation to illustrate the feasibility of the method.

2. MINIMIZATION OF Q

There are three reasonable approaches to finding the minimum. The first is the simplex method. It has been discussed in several other places; the original idea is by Nelder and Mead [4], and a comprehensive treatment can be found in the book by Walters, et al. [7]. The only input required is the function to be minimized and a starting value. It proceeds cautiously and slowly, but is almost always successful in finding the minimum. The second approach is to use a packaged minimization routine. Such routines sometimes require that partial derivatives of the function be available. The third approach is to obtain a set of equations by equating the partial derivatives to zero. The multi-variate version of the Newton-Raphson method could then be used to find the solution. When derivatives are needed they can be obtained by differentiating either Equation 1.2 or 1.5. The examples in this paper were done using the simplex method.

For the second and third approaches it is easy to write the partial derivative of Q.

$$\partial Q/\partial\theta_j = 2\sum_{i=1}^{k} w_i\left[G(c_i;\boldsymbol{\theta}) - G_n(c_i)\right]G^{(j)}(c_i;\boldsymbol{\theta}) \tag{2.1}$$

where the final factor ($G^{(j)}(c_i;\boldsymbol{\theta})$) is the partial derivative of the model functional with respect to θ_j. To simplify the notation, the model functional evaluated at c_i will be written G_i, the reference to $\boldsymbol{\theta}$ being implicit and the dependence on c_i being reflected by the subscript. Similarly, the empirical functional will be written $G_{n,i}$. Equations 1.1 and 2.1 become

$$Q = \sum_{i=1}^{k} w_i\,(G_i - G_{n,i})^2$$

and

$$\partial Q/\partial\theta_j = 2\sum_{i=1}^{k} w_i\,(G_i - G_{n,i})\,G_i^{(j)}\,. \tag{2.2}$$

3. STATISTICAL PROPERTIES OF MINIMUM DISTANCE ESTIMATES

The minimum distance estimate is an implicit function (as given in Equation 2.1) of $\boldsymbol{G}_n$, the vector of empirical functionals. The properties of such an estimator can be obtained by using Theorem 2 and Corollary 1 from Benichou and Gail [2]. The theorem requires that the estimator be an implicit function of random variables to which the Central Limit Theorem can be applied. This is true for both situations. The LEV is a sample average of independent observations and the empirical distribution function is a binomial proportion. We have

$$n^{1/2}(\boldsymbol{G}_n - \boldsymbol{\mu}) \to N(\boldsymbol{0}, \boldsymbol{\Sigma}) . \tag{3.1}$$

The i^{th} element of $\boldsymbol{\mu}$ is $\mu_i = \mathrm{E}(G_{n,i}) = G_i$ (at least for the two functionals used in this paper). Let the ij^{th} element of $\boldsymbol{\Sigma}$ be σ_{ij}.

The next item to be satisfied is that the k functions in Equation 2.1 have continuous first partial derivatives with respect to the elements of $\boldsymbol{\theta}$. These form a $p \times p$ matrix $\boldsymbol{A}$. The jl^{th} element is

$$a_{jl} = \partial^2 Q / \partial\theta_j \theta_l = 2 \sum_{i=1}^{k} w_i \, G_i^{(j)} G_i^{(l)} + 2 \sum_{i=1}^{k} w_i \, (G_i - G_{n,i}) \, G_i^{(j,l)} . \tag{3.2}$$

So, to satisfy the conditions of the theorem, the model functional must have continuous second partial derivatives with respect to the parameters. This is true for most distributions in common use for insurance losses. It is also necessary that $\boldsymbol{A}$ have a non-zero determinant when evaluated at the true parameter value. All that is necessary to complete this analysis is that it be non-zero at the estimated value of $\boldsymbol{\theta}$.

The next matrix, $\boldsymbol{B}$ $(p \times k)$, has jl^{th} element

$$b_{jl} = \partial^2 Q / \partial\theta_j \, G_{n,l} = -2 w_l \, G_l^{(j)} . \tag{3.3}$$

It is necessary that $\boldsymbol{A}^{-1}\boldsymbol{B}$ have at least one non-zero element.

The theorem then states that, as the sample size goes to infinity, there will be a unique solution, $\tilde{\boldsymbol{\theta}}$, to the equations and

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow N(\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{B\Sigma B}'\boldsymbol{A}^{-1}). \tag{3.4}$$

This verifies that the minimum distance estimator is consistent and asymptotically unbiased and, even though it is not likely to have minimum variance, at least we will be able to estimate the variance.

4. EXAMPLES

Example One

The first example consists of losses from the Insurance Services Office (ISO) increased limits project for general liability (Table 2) coverage. The accident year is 1986 and the losses are those reported at Lag 1. Actual losses are given in Table 1. This example uses fewer size-of-loss intervals. For simplification, the average loss in each interval was taken as the midpoint. One problem is the existence of multiple policy limits in the ISO data set. These are difficult to deal with as it is unlikely that actual losses can be determined for those cases that exceed the upper limit. There are two such cases in this data set. One loss is known to exceed \$25,000; the other exceeds \$500,000. The easiest reasonable way to adjust for this problem may be to replace these values with the conditional (on being above the upper limit) median (as the mean may not exist) from a rough estimate of the final model. For this illustration the values \$38,865 and \$769,061 were used. They were incorporated in the calculation of the empirical LEVs in Table 1.

For this illustration, the only distribution being considered is the Pareto distribution. ISO rejected it as a useful model (opting for a mixture of two Pareto distributions), but it will serve as a good example mostly because all the required derivatives are easy to compute. About the only other distributions that have this property are the lognormal and inverse Gaussian. Should analytical derivatives not be available, approximate differentiation must be employed. This example also proves to be somewhat simple, as there is no deductible involved. The relevant quantities for the Pareto distribution where $\boldsymbol{\theta} = (\alpha, \lambda)'$ are:

$$F(x;\boldsymbol{\theta}) = 1 - \left(\frac{\lambda}{\lambda + x}\right)^{\alpha}, x, \alpha, \lambda > 0 ,$$

$$L(c;\boldsymbol{\theta}) = \frac{\lambda}{\alpha - 1}\left[1 - \left(\frac{\lambda}{\lambda + c}\right)^{\alpha - 1}\right],$$

$$L^{(1)}(c;\boldsymbol{\theta}) = -\frac{L(c;\boldsymbol{\theta})}{\alpha - 1} - \frac{\lambda}{\alpha - 1}\left(\frac{\lambda}{\lambda + c}\right)^{\alpha - 1}\ln\left(\frac{\lambda}{\lambda + c}\right),$$

$$L^{(2)}(c;\boldsymbol{\theta}) = \frac{L(c;\boldsymbol{\theta})}{\lambda} - \frac{c\lambda^{\alpha - 1}}{(\lambda + c)^{\alpha}} . \tag{4.1}$$

Maximum likelihood estimation produced the estimates $\hat{\alpha} = 1.482595$ and $\hat{\lambda} = 705.785$. The estimated covariance matrix of these estimators is

$$\begin{bmatrix} 0.0020473 & 1.3680 \\ 1.3680 & 1{,}090.5 \end{bmatrix}.$$

Minimization of Q using weights of 1 at all endpoints (the value 10,000,000 was arbitrarily selected to replace ∞) produced the minimum LEV estimates of $\tilde{\alpha} = 1.3388257$ and $\tilde{\lambda} = 590.32670$. The value of Q at the minimum is 8,619 compared to a value of 196,244 using the maximum likelihood estimates (which were used as a starting point for the simplex method). Table 2 shows the LEVs for both maximum likelihood and minimum LEV estimation. The wide discrepancy between these two estimators may well indicate that the Pareto model is not suitable for these data.

TABLE 1

ISO Loss Data

Lower Limit	Upper Limit	Number of Losses	LEV (at upper limit)
$ 0	$ 50	482	$ 48.19
50	100	574	92.41
100	150	478	132.68
150	200	431	169.54
200	250	343	203.49
250	300	337	234.89
300	400	616	290.52
400	500	518	337.64
500	600	311	378.53
600	700	263	415.10
700	800	256	447.78
800	900	170	477.26
900	1,000	212	503.86
1,000	1,500	501	610.12
1,500	2,000	297	686.41
2,000	2,500	181	744.74
2,500	3,000	116	791.91
3,000	3,500	93	831.24
3,500	4,000	72	864.37
4,000	4,500	40	893.29
4,500	4,999	32	919.45
4,999	5,000	18	919.50
5,000	6,000	59	962.39
6,000	7,500	53	1,014.12
7,500	9,999	60	1,079.07
9,999	10,000	6	1,079.09
10,000	12,000	21	1,117.10
12,000	15,000	27	1,163.30
15,000	20,000	22	1,221.89
20,000	25,000	23	1,263.58
25,000	35,000	15	1,318.42
35,000	50,000	15	1,366.87
50,000	75,000	6	1,408.19
75,000	100,000	3	1,432.60
100,000	250,000	3	1,511.48
250,000	500,000	0	1,586.60
500,000	1,000,000	2	1,661.72
1,000,000	∞	0	1,661.72
Total		6,656	

TABLE 2

LEVs

Limit	Empirical LEV	MLE LEV	MinLEV
$ 50	$ 48.19	$ 47.58	$ 47.34
100	92.41	90.59	89.97
150	132.68	129.88	128.66
200	169.54	165.90	164.00
250	203.49	199.09	196.47
300	234.89	229.80	226.44
400	290.52	284.92	280.14
500	337.64	333.10	327.03
600	378.53	375.70	368.49
700	415.10	413.72	405.53
800	447.78	447.93	438.91
900	477.26	478.93	469.23
1,000	503.86	507.19	496.93
1,500	610.12	618.64	607.11
2,000	686.41	697.87	686.67
2,500	744.74	757.95	747.94
3,000	791.91	805.55	797.21
3,500	831.24	844.47	838.05
4,000	864.37	877.08	872.70
4,500	893.29	904.92	902.63
4,999	919.45	929.02	928.82
5,000	919.50	929.06	928.87
6,000	962.39	969.06	972.98
7,500	1,014.12	1,014.86	1,024.62
9,999	1,079.07	1,068.76	1,087.18
10,000	1,079.09	1,058.77	1,087.20
12,000	1,117.10	1,100.01	1,124.49
15,000	1,163.30	1,135.25	1,167.65
20,000	1,221.89	1,176.11	1,219.33
25,000	1,263.58	1,201.50	1,256.47
35,000	1,318.42	1,242.33	1,307.84
50,000	1,366.87	1,276.61	1,356.64
75,000	1,408.19	1,309.30	1,405.70
100,000	1,432.60	1,329.01	1,436.75
250,000	1,511.48	1,376.53	1,518.03
500,000	1,586.60	1,400.93	1,564.89
1,000,000	1,661.72	1,418.41	1,601.99
10,000,000	1,661.72	1,457.70	1,712.80

To estimate the asymptotic variance we need the variance of the empirical LEVs which are computed using:

$$\mathrm{E}[\min(X, c_i)^2] - \{\mathrm{E}[\min(X, c_i)]\}^2.$$

They are:

$$\sigma_{ii} = \mathrm{Var}\,(\min\,(X, c_i))$$

$$= \int_0^{c_i} x^2 f(x;\boldsymbol{\theta})dx + c_i^2\left[1 - F(c_i;\boldsymbol{\theta})\right] - L_i^2 = {}_2L_{ii} - L_i^2\,,$$

$$\sigma_{ij} = \mathrm{Cov}(\min\,(X, c_i), \min\,(X, c_j))$$

$$= \int_0^{c_i} x^2 f(x;\boldsymbol{\theta})dx + \int_{c_i}^{c_j} c_i\, x f(x;\boldsymbol{\theta})dx + c_i c_j\left[1 - F(c_j;\boldsymbol{\theta})\right] - L_i L_j$$

$$= {}_2L_{ij} - L_i L_j\,,\ \text{ for } i < j\,. \tag{4.2}$$

Note that if there is a deductible, d, the integral must start at d and the pdf and cumulative density function (cdf) must be modified to reflect the truncation.

For the Pareto distribution, with $i \leq j$,

$${}_2L_{ij} = \frac{2\lambda^2}{(\alpha-2)(\alpha-1)} - \frac{\lambda^\alpha(\lambda+c_i)^{-\alpha+1}(\alpha c_i + 2\lambda)}{(\alpha-2)(\alpha-1)} - \frac{\lambda^\alpha(\lambda+c_j)^{-\alpha+1}c_i}{(\alpha-1)}\,. \tag{4.3}$$

Using the 38 intervals and the estimated parameters produces a 38×38 matrix, which will not be presented here. The square root of the diagonal elements measures the standard deviation of the empirical LEVs based on a single observation. The standard deviation of the actual empirical LEVs can be estimated by dividing these values by the square root of the sample size (81.58). These standard deviations are presented for selected values in Table 3.

Calculation of the matrix $\boldsymbol{B}$ is relatively simple as Equation 3.3 requires only the first partial derivatives of the model LEVs. These were given in Equation 4.1. This matrix is not presented here.

Calculation of $\boldsymbol{A}$ requires the second partial derivatives of the model LEV. They are

$$L_i^{(1,1)} = -2\frac{L_i^{(1)}}{(\alpha-1)} - \frac{\lambda}{(\alpha-1)}\left(\frac{\lambda}{\lambda+c_i}\right)^{\alpha-1}\left[\ln\left(\frac{\lambda}{\lambda+c_i}\right)\right]^2,$$

$$L_i^{(1,2)} = L_i^{(2,1)} = \frac{L_i^{(1)}}{\lambda} - \frac{c_i}{\lambda+c_i}\left(\frac{\lambda}{\lambda+c_i}\right)^{\alpha-1}\ln\left(\frac{\lambda}{\lambda+c_i}\right),$$

$$L_i^{(2,2)} = \frac{L_i^{(2)}}{\lambda} - \frac{L_i}{\lambda^2} + \frac{c_i\lambda^{\alpha-2}(\lambda+c_i-\alpha c_i)}{(\lambda+c_i)^{\alpha+1}}. \tag{4.4}$$

For the data of the example, the matrix is

$$\boldsymbol{A} = \begin{bmatrix} 204{,}021{,}910 & -169{,}261.81 \\ -169{,}261.81 & 148.34278 \end{bmatrix}.$$

The estimated covariance matrix, $\boldsymbol{A}^{-1}\boldsymbol{B\Sigma B'A}^{-1}/6{,}656$ (the denominator is the sample size for this problem), is

$$\begin{bmatrix} 0.034751 & 33.571 \\ 33.571 & 32{,}765 \end{bmatrix}.$$

As expected, the minimum LEV estimator is inferior to maximum likelihood.

TABLE 3

STANDARD DEVIATIONS OF EMPIRICAL LEVs

Limit	LEV	Std. Dev.
$ 100	$ 92	0.3
250	203	1.0
500	338	2.3
1,000	504	4.6
2,500	745	9.9
5,000	920	15.7
10,000	1,079	23.2
25,000	1,264	36.0
50,000	1,367	48.3
100,000	1,433	63.3
500,000	1,587	113.4
1,000,000	1,662	144.2

Example Two

The second example concerns medical malpractice claim count development. The data are from Accomando and Weissner [1]. Cumulative numbers of claims were recorded at intervals of six months through 168 months. The data are presented in Table 4.

Maximum likelihood estimation revealed that the Burr distribution provides a good fit. The distribution function is

$$F(x) = \frac{1 - \left(\dfrac{\lambda^\tau}{\lambda^\tau + x^\tau}\right)^\alpha}{1 - \left(\dfrac{\lambda^\tau}{\lambda^\tau + 168^\tau}\right)^\alpha} . \tag{4.5}$$

TABLE 4

MEDICAL MALPRACTICE CLAIM COUNT DEVELOPMENT

Lag	Claims	Fn	F – MLE	F – Min$F$
6	4	0.0086	0.0020	0.0026
12	10	0.0216	0.0173	0.0194
18	18	0.0389	0.0574	0.0604
24	56	0.1210	0.1257	0.1276
30	101	0.2181	0.2142	0.2139
36	137	0.2959	0.3101	0.3079
42	199	0.4298	0.4025	0.3998
48	232	0.5011	0.4860	0.4838
54	261	0.5637	0.5585	0.5576
60	285	0.6156	0.6207	0.6212
66	307	0.6631	0.6736	0.6754
72	331	0.7149	0.7188	0.7216
78	352	0.7603	0.7574	0.7611
84	369	0.7970	0.7907	0.7949
90	380	0.8207	0.8195	0.8241
96	389	0.8402	0.8447	0.8493
102	396	0.8553	0.8668	0.8714
108	409	0.8834	0.8863	0.8907
114	414	0.8942	0.9036	0.9077
120	416	0.8985	0.9190	0.9229
126	423	0.9136	0.9329	0.9363
132	440	0.9503	0.9454	0.9484
138	445	0.9611	0.9567	0.9592
144	453	0.9784	0.9669	0.9690
150	455	0.9827	0.9763	0.9778
156	461	0.9957	0.9849	0.9859
162	463	1.0000	0.9927	0.9933
168	463	1.0000	1.0000	1.0000

The denominator is required to reflect the truncation of the data at 168 months. The maximum likelihood estimates of the parameters are $\hat{\alpha} = 0.40274$, $\hat{\lambda} = 34.224$, and $\hat{\tau} = 3.1181$. The values of $F(x)$ for this model are presented in Table 4.

The asymptotic covariance matrix of the maximum likelihood estimates is

$$\begin{bmatrix} 0.017336 & 0.57018 & -0.035566 \\ 0.57018 & 20.656 & -1.2135 \\ -0.035566 & -1.2135 & 0.10703 \end{bmatrix}.$$

For minimum distance estimation, the weights were selected as follows: if $F_{n,i} < 0.5$ the weight is 4, while if $F_{n,i} \geq 0.5$ the weight is $1/[F_{n,i}(1 - F_{n,i})]$. This places the smallest emphasis on the early durations and makes the weights proportional to the reciprocal of the variance at later durations (due to the omission of the sample size). Because the value of F_n at the last duration (162) is 1, the weight here is set equal to the one at duration 156. An alternative is to use the model distribution for the weights, changing them at each iteration as the parameters change. The minimum distance estimates are $\hat{\alpha} = 0.48798$, $\hat{\lambda} = 36.989$, and $\hat{\tau} = 2.9496$. These turn out to be very similar to the maximum likelihood estimates. A look at the distribution function in Table 4 verifies that this model does a better job of matching the distribution function, especially after the 95$^{\text{th}}$ percentile.

Estimation of the variance is messier than for the previous example due to the additional parameter and the complexity added by the denominator in Equation 4.5. For this illustration, the elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ were obtained by numerical differentiation. When this approximation was applied to the previous example, the answers matched to two significant digits. The elements of $\boldsymbol{\Sigma}$ are much easier to obtain. The ij^{th} element is

$$\sigma_{ij} = F_i(1 - F_j), \quad i \leq j\,. \tag{4.6}$$

The estimated covariance matrix is

$$\begin{bmatrix} 0.081077 & 2.6655 & -0.16625 \\ 2.6655 & 89.507 & -5.5313 \\ -0.16625 & -5.5313 & 0.33525 \end{bmatrix}.$$

This is about four to five times greater than the variances for the maximum likelihood estimate.

The goal of this application is to forecast the number of claims that will be reported after Lag 168. Using the Burr distribution it can be estimated as

$$\hat{\rho} = 463[1/F(168;\hat{\boldsymbol{\theta}}) - 1] = \frac{463}{\left[1 + \left(\frac{168}{\lambda}\right)^{\tau}\right]^{\alpha} - 1}, \tag{4.7}$$

where F is the untruncated Burr distribution. Inserting the maximum likelihood estimates yields $\hat{\rho} = 72.3998$, while doing the same for the minimum distance estimates yields $\hat{\rho} = 58.7556$. An estimate of the variance of these estimators can be obtained by finding the vector of partial derivatives (with respect to the parameters) of $\hat{\rho}$, $\boldsymbol{\delta}$, and then computing $\boldsymbol{\delta}\,\boldsymbol{\Sigma}\boldsymbol{\delta}$ where $\boldsymbol{\Sigma}$ is the covariance matrix of the parameter estimates. For the maximum likelihood estimate, the variance is 60.703 while for the minimum distance estimate it is 103.09. In the latter case, we can be about 95% confident that there are between 39 and 79 unreported claims.

5. A GOODNESS-OF-FIT TEST

If the model selected is correct, the empirical $G_{n,i}$ will have an approximate multivariate normal distribution with a mean equal to the model G and a covariance matrix given by $\boldsymbol{\Sigma}/n$. If the true parameters were known,

$$n(\boldsymbol{G}_n - \boldsymbol{G})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{G}_n - \boldsymbol{G}), \tag{5.1}$$

where G is the vector of functionals at the true parameter value, would have a chi-square distribution with k degrees of freedom. With the parameters being estimated, it is not so clear what to do. The remainder of this section addresses that problem. The approach used here is similar to the one used to derive the distribution of the chi-square goodness-of-fit test statistic. An excellent exposition can be found in Moore [3]. It is based on the work of Rao [5].

Let $\boldsymbol{V}_n(\boldsymbol{\theta})$ be a $k \times 1$ vector with i^{th} element $w_i^{1/2}\,[G_n(c_i) - G(c_i \mid \boldsymbol{\theta})]$ so $Q = \boldsymbol{V}_n'(\boldsymbol{\theta})\boldsymbol{V}_n(\boldsymbol{\theta})$. Let $\boldsymbol{\theta}_0$ be the true parameter value and $\boldsymbol{R}$ be a $k \times p$ matrix with ij^{th} element

$$r_{ij} = w_i^{1/2}\frac{\partial G(c_i \mid \boldsymbol{\theta})}{\partial \theta_j}\bigg|_{\theta_j = \theta_{0j}}. \tag{5.2}$$

From Equation 2.2, we have

$$\sum_{i=1}^{k} w_i^{1/2}\Big[\, G_n(c_i) - G(c_i \mid \tilde{\boldsymbol{\theta}})\,\Big]\tilde{r}_{ij} = 0, \quad j = 1, \ldots, p \tag{5.3}$$

where $\tilde{r}_{ij}$ is r_{ij} except with the derivative evaluated at $\tilde{\theta}_j$. Next write

$$G_n(c_i) - G(c_i \mid \tilde{\boldsymbol{\theta}}) = G_n(c_i) - G(c_i \mid \boldsymbol{\theta}_0) + G(c_i \mid \boldsymbol{\theta}_0) - G(c_i \mid \tilde{\boldsymbol{\theta}}) \tag{5.4}$$

$$= G_n(c_i) - G(c_i \mid \boldsymbol{\theta}_0) - \sum_{j=1}^{p}\Big[\, w_i^{-1/2} r_{ij} + o_p(1)\,\Big](\tilde{\theta}_j - \theta_{0j})$$

using a Taylor series approximation. Multiplying both sides by $w_i^{1/2}$ and arranging the elements in a $k \times 1$ vector produces

$$\boldsymbol{V}_n(\tilde{\boldsymbol{\theta}}) = \boldsymbol{V}_n(\boldsymbol{\theta}_0) - \boldsymbol{R}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - o_p(1)\,(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\,. \tag{5.5}$$

Assuming continuity of the elements of $\boldsymbol{R}$ as a function of $\boldsymbol{\theta}$, $\tilde{r}_{ij} = r_{ij} + o_p(1)$. Substituting this and Equation 5.5 into Equation 5.3 gives

$$\boldsymbol{0} = (\boldsymbol{R}' + o_p(1))\boldsymbol{V}_n(\tilde{\boldsymbol{\theta}})$$

$$= (\boldsymbol{R}' + o_p(1))\,[\boldsymbol{V}_n(\boldsymbol{\theta}_0) - \boldsymbol{R}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - o_p(1)\,(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)]$$

$$= \boldsymbol{R}'V_n(\boldsymbol{\theta}_0) - \boldsymbol{R}'\boldsymbol{R}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1) . \quad (5.6)$$

Rearranging gives

$$\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = (\boldsymbol{R}'\boldsymbol{R})^{-1}\, \boldsymbol{R}'V_n(\boldsymbol{\theta}_0) + o_p(1) . \quad (5.7)$$

Substituting Equation 5.7 into Equation 5.5 yields

$$V_n(\tilde{\boldsymbol{\theta}}) = V_n(\boldsymbol{\theta}_0) - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\, \boldsymbol{R}'V_n(\boldsymbol{\theta}_0) + o_p(1)$$

$$= [\boldsymbol{I} - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\, \boldsymbol{R}']V_n(\boldsymbol{\theta}_0) + o_p(1)$$

$$= \boldsymbol{C}V_n(\boldsymbol{\theta}_0) + o_p(1) . \quad (5.8)$$

Note that $\boldsymbol{C}$ is idempotent and assume that it is of rank $k - p$, as will most certainly be the case. Next observe that

$$V_n(\boldsymbol{\theta}_0) \sim N(\boldsymbol{0}, n^{-1}\boldsymbol{W}^{1/2}\boldsymbol{\Sigma}\boldsymbol{W}^{1/2}) , \quad (5.9)$$

where $\boldsymbol{W}^{1/2}$ is diagonal with i^{th} diagonal element $w_i^{1/2}$ and $\boldsymbol{\Sigma}$ is as in Equation 3.1. Therefore

$$V_n(\tilde{\boldsymbol{\theta}}) \sim N(\boldsymbol{0}, \boldsymbol{S} = n^{-1}\boldsymbol{C}\boldsymbol{W}^{1/2}\boldsymbol{\Sigma}\boldsymbol{W}^{1/2}\boldsymbol{C}) . \quad (5.10)$$

In general, if $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{S})$ then $\boldsymbol{X}'\, \boldsymbol{S}^-\, \boldsymbol{X} \sim \chi^2(m)$ where m is the rank of $\boldsymbol{S}$ and $\boldsymbol{S}^-$ is a generalized inverse (Moore [3], Theorem 2). One definition (among many that are equivalent) of a generalized inverse is that $x = \boldsymbol{S}^-y$ solves $y = \boldsymbol{S}x$ provided y is in the column space of $\boldsymbol{S}$. That is, if there is a solution to the equation, then $\boldsymbol{S}^-$ will provide it. A discussion of generalized inverses can be found in Searle [6]. At first it appears that this test is arbitrary, because the generalized inverse is not unique. But, for $\boldsymbol{X}$ in the column space of $\boldsymbol{S}$, $\boldsymbol{X}'\,\boldsymbol{S}^-\boldsymbol{X}$ will take on the same value, regardless of the form of the generalized inverse selected. For the normal distribution, the probability that this will happen is 1. Because the normal distribution in Equation 5.10 is approximate, it is possible that in practice, the value will depend slightly on the form of the generalized inverse selected. The test statistic is then

$$V_n(\tilde{\theta})'S^- V_n(\tilde{\theta}) = (G_n - G)'W^{1/2}S^- W^{1/2}(G_n - G), \qquad (5.11)$$

which is very similar to Equation 5.1.

For the second example, the value using Equation 5.1 is 70.53; it is 70.115 using Equation 5.11 with the Moore-Penrose generalized inverse. Both values clearly exceed the 5% critical value for 24 degrees of freedom (36.42).

This test indicates that a better choice of weights would have been appropriate. One such choice, from pure statistical (as opposed to actuarial) considerations, would be the reciprocals of the diagonal elements of $\boldsymbol{\Sigma}$. Aside from being an advance attempt to pass the hypothesis test, it makes sense in that the expected value of each term of $\boldsymbol{Q}$ is $1/n$. Thus, each term is making an approximately equal contribution to the criterion. For the Pareto example, a look at Table 3 shows that the weights would be decreasing with c_i. Again, this makes statistical sense, as for low limits virtually any reasonable model will produce an LEV that is just a little bit below c_i, and the empirical LEV will also be in that range. At the larger limits, there is likely to be much more sampling error and, therefore, wider variations should be tolerated. However, for actuarial purposes, one might come to the opposite conclusion. Once put to use, the model will be evaluated only at the larger limits, and so it is there where deviations from the sample should be small.

A more direct form of hypothesis test would be one based on $\boldsymbol{Q}$. This would be similar to the Cramer-von Mises test for comparing a model cdf to the empirical cdf. It has the advantage of being independent of the weights in the sense that the parameter estimate is, by definition, the one that minimizes the test statistic. However, this involves extra work as the distribution of $\boldsymbol{Q}$ under the null hypothesis is not so easy to obtain and depends heavily on the unknown $\boldsymbol{\theta}$.

6. SIMULATION

The theorem and hypothesis test are both asymptotic results. Also, both employ the replacement of the true parameter value by the estimate to complete the calculations. In this section, a simulation study is conducted to provide some feel for the accuracy of the method.

The true model selected for the study is Pareto with $\alpha = 3$ and $\lambda = 500$. The empirical LEV is obtained at 12 points: 20, 40, 65, 90, 130, 180, 250, 350, 575, 850, 1,300, and 2,000. At each simulation, 500 observations were generated. The parameters are then estimated by the minimum LEV method using weights of 1, 1, 1, 1, 1, 1, 1, 2, 4, 8, 16, and 16. The covariance matrix was also estimated, using Equation 3.4. Finally, the chi-square goodness-of-fit test statistic was computed using both Equations 5.1 and 5.11. The latter was done with two different algorithms for the generalized inverse, the Moore-Penrose and a sweep method. If the results in Sections 3 and 5 hold, the following should be observed:

1. The sample mean of the parameter estimates should be close to the true value. This will indicate that the estimator is unbiased.
2. The sample covariance matrix of the parameter estimates should be close to the matrix given by Equation 3.4 using the true parameter values. This will indicate that the theorem gives reasonable results for samples of size 500.
3. The estimated covariance matrices should have an average that is close to the matrix given by Equation 3.4 using the true parameter values. This will indicate that the replacement of the true values by the estimates does not distort the covariance estimation (on average).
4. The goodness-of-fit test statistics should have a sample mean of 10 and a sample variance of 20. This will indicate that the chi-square distribution with 10 degrees is reasonable. Also, 95% of the time the test statistic should be less than 18.307, and 99% of the time it should be less than 23.209. This will confirm that the significance level is as advertised.

A run of 1,000 simulations was conducted. The asymptotic covariance matrix for maximum likelihood estimation is

$$\begin{bmatrix} 0.3217 & 66.89 \\ 66.89 & 14{,}750 \end{bmatrix}.$$

The asymptotic covariance matrix for minimum LEV estimation is

$$\begin{bmatrix} 0.6640 & 120.3 \\ 120.3 & 21{,}830 \end{bmatrix}.$$

The sample means of the minimum LEV estimates were 3.161 for α and 535.1 for λ. The standard errors for α and λ are 0.023 and 4.9, respectively, indicating that, for a sample size of 500, there is bias in these estimates. The sample variances were 0.5133 for α and 24,150 for λ, and the sample covariance was 108.8. These are close to those given by the asymptotic approximation, indicating that Point 2 holds for this problem. With both estimates having a positive bias, there is some cancellation of error. For example, the true mean is $500/2 = 250$ while the mean of the Pareto distribution using the sample means is $535.1/2.161 = 247.62$. Using the approximation for the covariance matrix yielded average variances of 1.178 and 54,196. These considerably overstate the true values, and so Point 3 does not hold. Finally, the basic chi-square test (Equation 5.1) accepted the model 94.8% of the time when a 5% significance level was used and 99.3% of the time when a 1% level was used. Using Equation 5.11 with the Moore-Penrose inverse yielded acceptance rates of 95.5% and 99.4%, while the sweep inverse accepted the model 95.4% and 99.4% of the time. Another indication of accuracy is the mean and variance of the chi-square statistics. They were 10.002 and 19.542 for Equation 5.1, 9.843 and 18.849 for the Moore-Penrose inverse, and 9.846 and 18.847 for the sweep inverse. Finally, the absolute differences in the chi-square statistics were averaged for each of the three possible comparisons. For Equation 5.1 versus Moore-Penrose, the average absolute difference was 0.158; and versus the sweep inverse, it was 0.162. The two versions of Equation 5.11 had an average absolute difference of 0.016. It appears that any of the three tests are likely to be valid.

REFERENCES

[1] Accomando, F., and Weissner, E., "Report Lag Distributions: Estimation and Application to IBNR Counts," *Transcripts of the 1988 Casualty Loss Reserve Seminar,* 1988, p. 1038.

[2] Benichou, J., and Gail, M., "A Delta Method for Implicitly Defined Random Variables," *The American Statistician,* Vol. 43, 1989, p. 41.

[3] Moore, D., "Chi-Square Tests," in *Studies in Statistics,* R.V. Hogg, ed., Mathematical Association of America, 1978, p. 66.

[4] Nelder, J., and Mead, R., "A Simplex Method for Function Minimization," *The Computer Journal,* Vol. 7, 1965, p. 308.

[5] Rao, C.R., "Theory of the Method of Estimation by Minimum Chi-Square," *Bulletin of the International Statistics Institute,* Vol. 35, 1955, p. 25.

[6] Searle, S., *Linear Models,* New York, Wiley, 1971.

[7] Walters, F., Parker, L., Morgan, S., and Deming, S., *Sequential Simplex Optimization,* Boca Raton, Florida, CRC Press, 1991.