Volume LXV, Part 1 No. 123

PROCEEDINGS

May 21, 22, 23, 24, 1978

ESTIMATION OF THE DISTRIBUTION OF REPORT LAGS BY THE METHOD OF MAXIMUM LIKELIHOOD

EDWARD W. WEISSNER

Often when we are pricing an insurance contract or setting an IBNR reserve, it would be very useful to know the underlying distribution of the time delay between the time a claim occurs and the time the claim is reported. The purpose of this paper is to estimate this distribution. Specifically, we introduce a procedure, based on the method of maximum likelihood, which can be used on immature claims data to estimate the distribution of the time delay between the time a claim occurs and the time the claim is reported.

We shall refer to this time delay, the elapsed time between the time of occurrence and the time the insurer records it on its books, as a *report lag*. While the distribution of these report lags would most likely be unknown, one might, based on experience and knowledge, be willing to assume that the underlying distribution is Poisson, exponential, log-normal or some other well known probability law. Further, if a random sample of report lags were available, one could use some statistical estimation procedure (i.e., maximum likelihood) to estimate the unknown parameters of the assumed distribution. Thus, a good estimate of the report lags distribution would be available.

Unfortunately, however, a random sample of current report lags is not usually available, especially for some of the long-tail casualty sublines like medical malpractice. We do have for each accident period, however, a cumulative record of the number of claims received over time. Table I, using accident month, is typical (though abbreviated for convenience).

TABLE I

Total Claims Reported

Accident Month	Report Month Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
3 Mar	8	13	21	25	31	37	42	44	44	45
4 Apr		5	15	22	32	34	37	39	42	43
5 May			4	20	27	35	40	43	45	47
6 June				4	14	25	31	36	39	41
7 July					7	15	24	29	33	43
8 Aug						4	22	29	35	42
9 Sept							5	15	28	36
10 Oct								11	22	31
11 Nov									10	25
12 Dec										8
									(Total)	361

This table can be restructured to yield the number of claims received during each report period for each accident period. If we assume that all claims occur and are reported at the middle of a period, then the restructured table also yields for each accident period the frequency of various report lags. Table II is the restructured Table I.

TABLE II

Number of Claims with a Report Lag of K Months

Accident Month	K = 0	1	2	3	4	5	6	7	8	9
3 Mar	8	5	8	4	6	6	5	2	0	1
4 Apr	5	10	7	10	2	3	2	3	1	
5 May	4	16	7	8	5	3	2	2		
6 June	4	10	11	6	5	3	2			
7 July	7	8	9	5	4	10				
8 Aug	4	18	7	6	7					
9 Sept	5	10	13	8						
10 Oct	11	11	9							
11 Nov	10	15								
12 Dec	8									

Hence, referring to Tables I and II, we observe that for the March accident month, 21 claims were reported by the end of May, 8 claims were reported in May, and therefore 8 claims have a report lag of 2 months. If we could assume that the 361 report lags from all the accident periods were a random sample, we could proceed as above. However, this sample of report lags is incomplete, immature, and biased toward small report lags. All the unreported claims in any accident period will yield only larger report lags. Hence, we do not have a random sample.

We now present a procedure which may be used to estimate the complete distribution of report lags, given the above data.

1) To begin, let us consider only the March accident month data received through the end of December (see Tables I and II). Let n be the number of reported claims; here $n=45$. Let the 45 report lags be $x_1, x_2, \ldots, x_{45}$; here $x_1 = x_2 = \ldots = x_8 = 0$, $x_9 = x_{10} = \ldots = x_{13} = 1, \ldots,$ $x_{43} = x_{44} = 7$, and $x_{45} = 9$. Now, for the moment, assume that the underlying report lag distribution is exponential with parameter θ, unknown. Then the report lag density is given by

$$f(x/\theta) = \theta \cdot exp(-\theta x) \qquad 0 < x < \infty,$$
$$= 0 \qquad \text{otherwise.}$$

If *all* the March accident month occurrences were known to have been reported by December 31 (i.e., no unreported claims), then the sample of 45 report lags would clearly be a random sample from the above exponential law. However, we don't know that this set of report lags is complete; several claims may be unreported as yet. We have observed only the claims reported through the end of December, that is, the claims with report lag less than or equal to 9 months. (Since our data is rounded to the nearest month and the model is continuous, we have effectively observed all the claims with report lag less than or equal to 9.5 months.) Let c be the maximum possible report lag (plus .5) for the accident period; here $c = 9.5$. *While these observed report lags are not a random sample from the exponential law, they do constitute a random sample from an exponential law conditioned (truncated) to allow only report lags of 9.5 months or fewer.*

Since according to our exponential model (recall $c = 9.5$)

$$P[\text{report lag} \leq c] = \int_0^c f(x/\theta)\, dx = 1 - exp(-\theta c),$$

the conditional (truncated) report lag density, $f(x/\theta,c)$ is given by

$$\begin{aligned} f(x/\theta,c) &= f(x/\theta)/P[\text{report lag} \leqslant c] \\ &= [\theta \cdot exp(-\theta x)]/[1 - exp(-\theta c)] \end{aligned} \tag{1.1}$$

for $0 < x < c$; 0 otherwise. Let us now use the concept of maximum likelihood estimation to estimate θ.[1] The likelihood function for θ for the March accident month, $L(\theta)$, is given by (recall that $n = 45$, $c = 9.5$, and the x_i's are known)

$$\begin{aligned} L(\theta) &= L(\theta; x_1, x_2, \ldots x_n) \\ &= \Pi_j f(x_j/\theta, c) \\ &= [\theta^n \cdot exp(-\theta \cdot \Sigma_j x_j)]/[1 - exp(-\theta c)]^n. \end{aligned}$$

Taking natural logs, we obtain

$$ln\, L(\theta) = n \cdot ln\theta - \theta \cdot \Sigma_j x_j - n \cdot ln[1 - exp(-\theta c)].$$

It follows that

$$\begin{aligned} \frac{d\, ln\, L(\theta)}{d\theta} &= \frac{n}{\theta} - \Sigma_j x_j - \frac{n \cdot c \cdot exp(-\theta c)}{[1 - exp(-\theta c)]} \\ &= g(\theta). \end{aligned} \tag{1.2}$$

Let the right hand side of (1.2) be $g(\theta)$. The maximum likelihood estimate of θ is the value $\hat{\theta}$ for which $g(\hat{\theta}) = 0$.

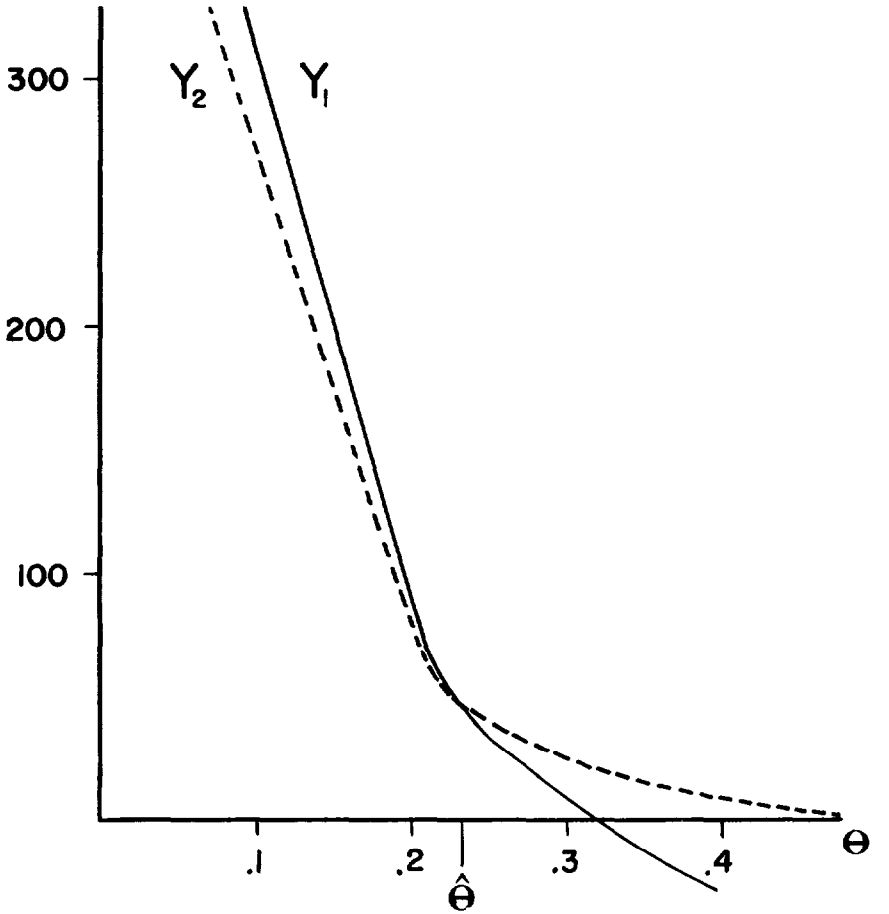
In our example then, we must solve

$$\frac{45}{\theta} - 140 - \frac{45(9.5)\, exp(-9.5\,\theta)}{1 - exp(-9.5\,\theta)} = 0$$

since $\Sigma x_j = 140$ (see Table II). To solve for θ, we might observe that the curves

$$\begin{aligned} y_1 &= (45/\theta) - 140 \\ y_2 &= 45(9.5)\, exp(-9.5\,\theta)/[1 - exp(-9.5\,\theta)] \end{aligned}$$

intersect when $\theta = \hat{\theta}$ (see Figure) and use this to determine θ.



Another approach would be to use a Newton-Raphson iteration to solve $g(\theta) = 0$. Since

$$g'(\theta) = -\frac{n}{\theta^2} + \frac{n \cdot c^2 \cdot exp(-\theta c)}{[1 - exp(-\theta c)]^2}$$

the Newton-Raphson iteration[2] for $\hat{\theta}$ is given by

$$\theta_{m+1} = \theta_m - g(\theta_m)/g'(\theta_m)$$

$$= \theta_m - \frac{(n/\theta_m) - \Sigma x_j - n \cdot c \cdot exp(-\theta_m c)/[1 - exp(-\theta_m c)]}{(-n/\theta_m^2) + n \cdot c^2 \cdot exp(-\theta_m c)/[1 - exp(-\theta_m c)]^2}$$

For our example, this reduces to

$$\theta_{m+1} = \theta_m - \frac{(45/\theta_m) - 140 - 45(9.5)\exp(-9.5\,\theta_m)/[1 - \exp(-9.5\,\theta_m)]}{(-45/\theta_m^2) + 45(9.5)^2\exp(-9.5\,\theta_m)/[1 - \exp(-9.5\,\theta_m)]^2}$$

This iteration is easy to program in APL on a mini-computer. Using a seed of $\theta_1 = .2$ (recall the mean of an exponential law is θ^{-1}; we thought it might be 5 months), we found

$$\theta_2 = .23442$$

$$\theta_3 = .23547$$

$$\theta_4 = .23547.$$

Hence, the maximum likelihood estimate of θ, using the March accident month data only, is $\hat{\theta} = .23547$. Thus, if you believe an exponential model for report lags is appropriate, you would use the exponential law with $\theta = .23547$ (and mean = 4.25 months).

Note that this value of θ is the parameter of the complete exponential report lags distribution as well as the parameter of the truncated exponential report lags distribution. Hence, the procedure, based on truncated distributions, yields an estimate of the complete report lags distribution.

As an example, if you would like to estimate the proportion of occurrences in any accident month which will have a report lag of at least 12 months, say, then the proportion p is given by

$$\begin{aligned} p &= P[\text{report lag of at least 12 months}] \\ &= P[lag \geq 11.5] \\ &= \int_{11.5}^{\infty} f(x/\theta = .23547)\,dx \\ &= \exp(-.23547(11.5)) \\ &= .067. \end{aligned}$$

(The shift from 12 to 11.5 is due to our correction for rounding.)

Or suppose you would like to estimate the number of unreported claims in the March accident month as of 12/31. Using an analysis similar to the above, we find that the proportion of occurrences reported within 9 months (use 9.5) is .893. If N is the total number of March accident month occurrences, then $.893N$ is the expected number of reported claims as of 12/31. Since the actual number of reported claims is 45 (see Table I), an estimate of N is found by solving $.893N = 45$. Thus for the March accident month, N is 50 which implies that the IBNR as of 12/31 is 5 claims.

2) Let us now use all of the available information to help us estimate θ. Let $n_3, n_4, \ldots, n_{12}$ be the respective numbers of reported claims through the end of December for the accident months March (3) through December (12). Then $n_3 = 45$, $n_4 = 43$, ..., and $n_{12} = 8$. Let $c_3, c_4, \ldots, c_{12}$ be the respective maximum possible report lag (plus .5) for the accident months March through December. It follows that $c_3 = 9.5$, $c_4 = 8.5$, ..., and $c_{12} = .5$ Finally, let x_{ij} be the j^{th} report lag in the i^{th} accident month ($i = 3,4,\ldots,12$ and $j = 1,2,\ldots,n_i$).

Then, as before, for the i^{th} accident month, the sample of report lags $\{x_{i1}, x_{i2}, \ldots, x_{in_i}\}$ obeys $f(x/\theta, c_i)$, the truncated exponential for the i^{th} accident month (see 1.1). Assuming that the accident months are independent, the generalized likelihood function for all the data, $L^*(\theta)$, is given by (see 1.1)

$$
\begin{aligned}
L^*(\theta) &= L^*(\theta; \text{all } x_{ij}\text{'s}) \\
&= \left[\prod_{j=1}^{n_3} f(x_{3j}/\theta, c_3)\right] \left[\prod_{j=1}^{n_4} f(x_{4j}/\theta, c_4)\right] \cdots \left[\prod_{j=1}^{n_{12}} f(x_{12j}/\theta, c_{12})\right] \\
&= \left\{exp(-\theta \cdot \Sigma_{ij} x_{ij}) \Big/ \prod_i [1 - exp(-\theta c_i)]^{n_i}\right\} \cdot \theta^{\Sigma_i n_i} \cdot
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\frac{d \ln L^*(\theta)}{d\theta} &= \frac{\Sigma_i n_i}{\theta} - \Sigma_{ij} x_{ij} - \Sigma_i \left\{\frac{n_i \cdot c_i \cdot exp(-\theta c_i)}{[1 - exp(-\theta c_i)]}\right\} \\
&= g^*(\theta) \cdot \qquad (2.1)
\end{aligned}
$$

Again, let the right hand side of (2.1) be $g^*(\theta)$. The maximum likelihood estimate of θ is the value $\hat{\theta}$ for which $g^*(\hat{\theta}) = 0$. In our example it means solving

$$\frac{361}{\theta} - 759 - \left\{ \frac{45(9.5)\exp(-9.5\,\theta)}{[1-\exp(-9.5\,\theta)]} + \frac{43(8.5)\exp(-8.5\,\theta)}{[1-\exp(-8.5\,\theta)]} + \cdots \right\} = 0$$

since $\Sigma n_i = 361$ and $\Sigma x_{ij} = 759$ (see Table II). Again, a Newton-Raphson iteration can be applied to solve $g(\theta) = 0$. If you do so and let $\theta_1 = .2$ again, then

$$\theta_2 = .24829$$

$$\theta_3 = .24971$$

$$\theta_4 = .24971.$$

Hence, using all the data through December 31, we conclude, for this exponential model, that the maximum likelihood estimate of θ is $.24971$. This implies that the average report lag is 4.00 months.

Again recall that this value of θ is the parameter of the complete exponential report lags distribution as well as a parameter in each of the truncated exponential report lags distributions. Thus, this procedure, based on truncated distributions, yields an estimate of the complete report lags distributions. Moreover, it also yields therefore an estimate of the complete average report lag.

According to the above analysis, the average report lag for all occurrences is 4 months. That is, when all the occurrences from a specific accident period have been reported, we expect that the average report lag will be 4 months.

Finally, based on the estimated average report lag of 4 months, we can demonstrate the accuracy of this procedure. The data in Table II was randomly generated using an exponential report lag with a mean of 4 months ($\theta = .250$) and increasing numbers of occurrences each accident month. While this data is therefore highly regular, we have obtained similar results on actual reinsurance claims data.

3) You need not of course assume an exponential model for the distribution of report lags or even a continuous model. This procedure however is easier to carry out for some models than for others. If you believe for instance that the model is log-normal, then the report lag density, $f(x/\mu, \sigma^2)$ is given by

$$f(x/\mu, \sigma^2) = (1/\sqrt{2\pi}\,\sigma \cdot x)\exp\left\{-.5[(\ln x - \mu)/\sigma]^2\right\}$$

and the truncated density is given by

$$f(x/\mu, \sigma^2, c) = f(x/\mu, \sigma^2)/\phi((\ln x - \mu)/\sigma)$$

where ϕ is the cumulative distribution function of the standard normal, $N(0,1)$. The procedure outlined yields the following equations for the joint maximum likelihood estimation of μ and σ^2:

$$\mu = \frac{\sum_{ij} \ln x_{ij}}{\sum_i n_i} + \sum_i \left\{ \frac{\sigma \cdot n_i}{\sum_i n_i} \right\} \left\{ \frac{\phi((\ln c_i - \mu)/\sigma)}{\phi((\ln c_i - \mu)/\sigma)} \right\}$$

$$\sigma^2 = \frac{\sum_{ij} (\ln x_{ij} - \mu)^2}{\sum_i n_i} + \sum_i \left\{ \frac{\sigma^2 \cdot n_i}{\sum_i n_i} \right\} \left\{ \frac{\phi((\ln c_i - \mu)/\sigma) \cdot ((\ln c_i - \mu)/\sigma)}{\phi((\ln c_i - \mu)/\sigma)} \right\}$$

where ϕ is the density of the standard normal, $N(0,1)$. To solve these equations for μ *and* σ^2, one could use successive substitution or a 2-dimensional Newton-Raphson iteration.[3] The Newton-Raphson method is much quicker!

4) We close with some procedural remarks. If after estimating the parameters of your model you wish to compare the model distribution and the observed sample distribution for an accident period, remember to use the truncated model distribution in your comparison. Secondly, it is important that the length of the accident (report) periods be relatively short (i.e., month or week). The report lag as defined can differ from the actual report lag by as much as one report period. For example, if an accident occurs on January 1 and is reported on March 31, the report lag based on the mid-points of the reporting months is 2 months, whereas the actual lag is 3 months. Thus, the shorter the period, the more precise the report lag is, the closer the data is to reality, and the better the estimation procedure works. Thirdly, it appears that the procedure works very well if there is at least one accident period with some "tail" lags to help give the early lags the appropriate balance. Finally, this kind of estimation using truncated distributions can also be useful in pricing problems where the losses are restricted only to large claims, only to small claims, or only to claims in a certain layer.[4]

---

[1] R. V. Hogg and A. T. Craig. *Introduction to Mathematical Statistics*. (3rd ed.) MacMillan, New York, 1970, p. 254.

[2] S. G. Kellison. *Fundamentals of Numerical Analysis*. Irwin, Homewood, Ill., 1975, p. 263.

[3] S. D. Conte and C. de Boor. *Elementary Numerical Analysis: An Algorithmic Approach*. (2nd ed.) McGraw-Hill, New York, 1972, p. 84.

[4] H. G. Verbeek. "An Approach to the Analysis of Claims Experience in Motor Liability Excess of Loss Reinsurance." *The ASTIN Bulletin*. Vol. 6, Part 3 (1972), pp. 195-202.