

Robust Confidence Intervals for Censored Data

Intervalli di Confidenza Robusti per Dati Censurati

Gianfranco Adimari

Dipartimento di Scienze Statistiche
Università di Padova
adimari@stat.unipd.it

Paolo Preo

Dipartimento di Scienze Statistiche
Università di Padova
paolo@stat.unipd.it

Riassunto: In questo lavoro si considera l'uso combinato di verosimiglianza empirica e equazioni di stima per dati censurati con l'obiettivo di ottenere intervalli di confidenza robusti per parametri e quantili di distribuzioni largamente usate nell'analisi di dati di durata. I risultati di alcuni esperimenti di simulazione relativi ad un'applicazione al modello log-normale mostrano l'utilità del metodo proposto.

Keywords: Empirical likelihood, Lifetime data, Location-scale model, Robustness.

1. Introduction

Let $y_1, \dots, y_n$ be n independent observations on a random variable Y with cumulative distribution function $F(y; \theta) = \text{pr}_\theta\{Y \leq y\}$ depending on an unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^s$, $s \geq 1$. Let $f(y; \theta)$ denote the probability density of Y and $S(y; \theta) = 1 - F(y; \theta)$ the survival function. A general M -estimator for θ is a solution of the equation

$$\sum_{i=1}^n \alpha(y_i; \theta) = 0, \quad (1)$$

where $\alpha(\cdot; \cdot) : \mathbb{R} \times \Theta \rightarrow \mathbb{R}^s$ is a suitable unbiased estimating function, i.e. $E\{\alpha(Y; \theta)\} = 0$. Examples include maximum likelihood estimators, generalized method of moment estimators, quasi-likelihood estimators and many robust estimators.

Robust techniques provide protection against outliers and model violation (see Hampel et al., 1986, as a general reference). Maximum likelihood estimators are efficient if the model is correctly specified but are typically not robust since they may be very sensitive to outliers. The level of the gross-error sensitivity specifies the extent to which a single outlier can influence an estimator and it is generally unbounded for the maximum likelihood estimator. Flexibility in choosing the shape of the function $\alpha(\cdot; \cdot)$ in (1) allows for the construction of estimates with desirable robustness properties. In particular, robust M -estimators with bounded gross-error sensitivity correspond to bounded estimating functions.

In lifetime data analysis, data are typically right censored by censoring values $t_1, \dots, t_n$ so that one observes $\{(z_i, \delta_i), i = 1, \dots, n\}$, where $z_i = \min(y_i, t_i)$ and $\delta_i = I(y_i \leq t_i)$, with $I(\cdot)$ the indicator function. When the censoring mechanism is *independent* (Kalbfleisch and Prentice, 1980, Section 5.2), as in many usual censoring schemes, the maximum likelihood estimate for θ is generally obtained by solving the likelihood equation

$$\sum_{i=1}^n \{\delta_i \Delta(z_i; \theta) + (1 - \delta_i) \Delta_c(z_i; \theta)\} = 0, \quad (2)$$

where $\Delta(y; \theta) = (\partial/\partial\theta) \log f(y; \theta)$ and $\Delta_c(y; \theta) = (\partial/\partial\theta) \log S(y; \theta)$.

Since $\Delta_c(z_i; \theta) = E\{\Delta(Y; \theta) | Y > z_i\}$ under mild regularity conditions, equation (2) suggests that the natural estimating equation analogue of (1) for censored data is

$$\sum_{i=1}^n \eta(z_i, \delta_i; \theta) = 0, \quad (3)$$

where $\eta(z_i, \delta_i; \theta) = \delta_i \alpha(z_i; \theta) + (1 - \delta_i) E\{\alpha(Y; \theta) | Y > z_i\}$. It follows that robust M -estimators for censored data can be readily obtained from robust estimating functions used in the uncensored case. James (1986) studies the asymptotic behaviour of the estimators obtained from (3), and we will refer such estimators as James-type M -estimators.

In this paper we combine empirical likelihood (see Owen, 2001, as a general reference) and estimating equations that define James-type M -estimators to obtain robust confidence intervals for parameters and quantiles of distributions used in lifetime data analysis.

2. The proposed method

For complete data, the combined use of empirical likelihood and estimating functions that define robust M -estimators for θ gives rise to confidence intervals that, in some way, maintain the robustness properties of the associated estimators. Typically such intervals have coverage level that is little sensitive with respect to small departures of the data distribution from the specified model. Furthermore, they are more accurate than those based on the asymptotic normality of the correspondent M -estimator.

In the censored case, if the objective is to construct confidence intervals with a certain level of protection against outlying observations or small deviations from the assumed model, we can opportunely combine empirical likelihood and estimating equations of the form (3), namely estimating equations that define James-type robust M -estimators. This approach is appealing at least for two reasons. First, it is quite general and apply to many practical situations where the censoring mechanism can be assumed to be independent. Second, unlike the normal approximation-based approach, the empirical likelihood one does not require estimation of the asymptotic variance of any statistic.

In the following, to illustrate the proposed method we consider inference for log-location-scale distributions, such as the log-normal and the Weibull distributions, which are frequently used in lifetime models in medical and biological sciences as well as in engineering. In lifetime data analysis, one often considers the logarithm of the lifetime variable so that the resulting transformed lifetime belongs to a location-scale family. The features of the location-scale family may facilitate the inference. From now on, we assume that the random variable Y describes the logarithm of lifetimes and that the z_i 's values are the transformed observed lifetimes.

Let $\theta = (\mu, \sigma)$. We focus on a location-scale family with densities $f(y; \mu, \sigma) = \sigma^{-1} f_*((y - \mu)/\sigma)$, where $f_*(\cdot)$ denotes the density of the standardized ($\mu = 0, \sigma = 1$) element of the family. For complete data, classical robust M -estimators for the location and scale parameters are defined through the estimating equation (1) with

$$\alpha(y; \theta) = \begin{pmatrix} \alpha_1(y; \mu, \sigma) \\ \alpha_2(y; \mu, \sigma) \end{pmatrix} = \begin{pmatrix} \alpha_1((y - \mu)/\sigma) \\ \alpha_2((y - \mu)/\sigma) \end{pmatrix} \quad (4)$$

for suitable (bounded) real-valued functions $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$. For censored data, related James-type location-scale estimators are obtained from (3) with

$$\eta(z, \delta; \theta) = \begin{pmatrix} \eta_1(z, \delta; \mu, \sigma) \\ \eta_2(z, \delta; \mu, \sigma) \end{pmatrix} = \begin{pmatrix} \delta\alpha_1((z - \mu)/\sigma) + (1 - \delta)\alpha_{1c}((z - \mu)/\sigma) \\ \delta\alpha_2((z - \mu)/\sigma) + (1 - \delta)\alpha_{2c}((z - \mu)/\sigma) \end{pmatrix}$$

where

$$\begin{aligned} \alpha_{1c}((z - \mu)/\sigma) &= E\{\alpha_1((Y - \mu)/\sigma) | Y > z\} \\ \alpha_{2c}((z - \mu)/\sigma) &= E\{\alpha_2((Y - \mu)/\sigma) | Y > z\}. \end{aligned} \quad (5)$$

The above expectations give the contribution of a censored observation to the estimating function η and, in general, must be evaluated numerically.

The empirical likelihood ratio (Owen, 2001, Sections 3.5 and 3.14) for (μ, σ) based on η and the sample $\{(z_i, \delta_i), i = 1, \dots, n\}$ is

$$R(\mu, \sigma) = \prod_{i=1}^n \{1 + \lambda' \eta(z_i, \delta_i; \mu, \sigma)\}^{-1},$$

where λ satisfies $\sum_{i=1}^n \eta(z_i, \delta_i; \mu, \sigma) \{1 + \lambda' \eta(z_i, \delta_i; \mu, \sigma)\}^{-1} = 0$. Profile empirical likelihood ratios for a single component or functions of (μ, σ) can be derived from $R(\mu, \sigma)$. For instance, the profile empirical likelihood ratio for the location parameter μ is

$$R(\mu) = \sup_{\sigma} R(\mu, \sigma).$$

On the other hand, if interest focuses on the q -th quantile τ of the distribution of Y , then we can use the relation $\tau = \mu + y_q^* \sigma$ which binds τ to the parameters (μ, σ) and the q -th quantile y_q^* of the standardized element of the location-scale family. Therefore, the empirical likelihood ratio for τ is

$$R(\tau) = \sup_{\sigma} R(\tau - y_q^* \sigma, \sigma).$$

Since $E\{\eta(Z, \delta; \mu, \sigma)\} = E\{\alpha(Y; \mu, \sigma)\} = 0$, the χ^2 approximation for the distribution of the empirical likelihood ratio statistic, $l(\cdot) = -2 \log R(\cdot)$, is justified under the same weak regularity conditions necessary in the uncensored case. When such conditions are satisfied, $\{\mu : l_E(\mu) \leq c_1(\gamma)\}$ and $\{\tau : l_E(\tau) \leq c_1(\gamma)\}$ are approximate confidence intervals for μ and τ respectively, with nominal coverage level $(1 - \gamma)$. Here $c_1(\gamma)$ is such that $\text{pr}\{\chi_1^2 > c_1(\gamma)\} = \gamma$.

3. Inference for the log-normal distribution

Now we consider inference for the log-normal distribution. In this case Y belongs to the normal family. For complete data, a well-known simultaneous location-scale robust estimator is the Huber estimator (Huber, 1964, Proposal 2) defined by (4) with $\alpha_1(u) = \psi(u; k_1)$ and $\alpha_2(u) = \psi^2(u; k_2) - a_{k_2}$ for appropriate constants k_1, k_2 and a_{k_2} . The function $\psi(u; k) = u \min\{1, k/|u|\}$ is the so-called Huber ψ -function.

Extension of the Huber estimator to the censored case is easy. In fact, for the normal model the expectations in (5) may be computed analytically. In particular, we have

$$\alpha_{1c}((z - \mu)/\sigma) = \begin{cases} k_1 & \text{if } (z - \mu)/\sigma > k_1 \\ \frac{k_1(1 - \Phi(k_1)) + \phi((z - \mu)/\sigma) - \phi(k_1)}{1 - \Phi((z - \mu)/\sigma)} & \text{if } |(z - \mu)/\sigma| \leq k_1 \\ \frac{k_1 \Phi((z - \mu)/\sigma)}{1 - \Phi((z - \mu)/\sigma)} & \text{if } (z - \mu)/\sigma < -k_1 \end{cases}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal and $\phi(\cdot)$ its density. Function $\alpha_{2c}((z - \mu)/\sigma)$ has a similar expression.

To show the usefulness of the proposed method, Table 1 gives some results of a Monte Carlo study designed to compare the empirical likelihood-based procedure (ELR) with the classical method based on the log-likelihood ratio statistic (LLR). Table 1 gives the empirical coverage probabilities of the $(1 - \gamma)$ -confidence intervals for the ninth decile of the distribution of the log lifetimes, for two sample sizes and different censoring rates (π). Each simulation experiment is based on 5000 trials. The log lifetimes are generated from the contamination model $F_\varepsilon = (1 - \varepsilon)\Phi + \varepsilon G$, where G denotes the cumulative distribution function of a normal with $\mu = 0$ and $\sigma = 5$. The contamination percentage ε is set at 0% (a) and at 5% (b). Type I (fixed) censoring is considered. The Huber estimator is used, with $k_1 = k_2 = 1.1$. Observe that a confidence interval for the ninth decile of the distribution of the lifetime variable is given by $\exp(\cdot)$ applied to each value in the confidence interval for the quantile of the distribution of the log lifetimes.

Table 1: Empirical coverage probabilities of the confidence intervals for the ninth decile of the distribution of the log lifetimes.

$1 - \gamma$		n=20		n=40		
		$\pi=0.2$	$\pi=0.4$	$\pi=0.2$	$\pi=0.4$	$\pi=0.6$
0.99	ELR	0.986	0.994	0.993	0.993	0.994
	LLR	0.986	0.990	0.990	0.990	0.992
0.95	ELR	0.952	0.962	0.952	0.957	0.962
	LLR	0.942	0.945	0.948	0.946	0.947
0.90	ELR	0.907	0.920	0.906	0.909	0.918
	LLR	0.895	0.893	0.899	0.891	0.895

(a)

$1 - \gamma$		n=20		n=40		
		$\pi=0.2$	$\pi=0.4$	$\pi=0.2$	$\pi=0.4$	$\pi=0.6$
0.99	ELR	0.989	0.991	0.989	0.993	0.994
	LLR	0.948	0.917	0.860	0.814	0.772
0.95	ELR	0.958	0.960	0.953	0.958	0.968
	LLR	0.870	0.854	0.765	0.718	0.688
0.90	ELR	0.913	0.911	0.902	0.915	0.923
	LLR	0.809	0.786	0.700	0.653	0.630

(b)

References

- Hampel, F. R., Ronchetti, E. M., Rosseeuw, P. S. and Stahel, W. A. (1986). *Robust Statistics: the Approach Based on Influence Function*, Wiley.
- Huber, P. J. (1964). Robust estimation of a location parameter, *Ann. Math. Stat.* **35**, 73–101.
- James, I. R. (1986). On estimating equations with censored data, *Biometrika* **73**, 35–42.
- Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley.
- Owen, A. B. (2001). *Empirical Likelihood*, Chapman and Hall.