Empirical likelihood analysis of the rank estimator for the censored accelerated failure time model

By MAI ZHOU

Department of Statistics, University of Kentucky, Lexington, KY, 40506 U.S.A.
mai@ms.uky.edu

Summary

We use the empirical likelihood method to derive a test and thus a confidence interval based on the rank estimators of the regression coefficient in the accelerated failure time model. Standard chi-squared distributions are used to calculate the p-value and to construct the confidence interval. Simulations and examples show that the chi-squared approximation to the distribution of the log empirical likelihood ratio performs well, and has some advantages over the existing methods.

Some key words: Gehan statistic; Survival data; Weighted log-rank statistic; Wilks theorem.

1. Introduction

The semiparametric accelerated failure time model (Wei, 1992) is a linear regression model in which the responses are the logarithms of survival times and the error term distribution is unspecified. It provides a useful alternative model to the Cox proportional hazards model for analysing censored survival data. Starting with Prentice (1978), many people have studied a rank-based estimation method in the accelerated failure time model with censored data, including Tsiatis (1990), Wei et al. (1990), Ritov (1991), Lai & Ying (1991) and Ying (1993). A nice summary can be found in Chapter 7 of Kalbfleisch & Prentice (2002).

A large-sample study of the rank-based estimator is given by Lai & Ying (1991), Ying (1993) and several subsequent papers including Jin et al. (2003) which deals with the weighted version of the rank-based estimator. They show that the estimator is asymptotically normally distributed but the asymptotic variance of the estimator involves the hazard function and the derivative of the hazard function of the unknown error term. The estimation of such functions can be highly unstable. To make things worse, the number

of subjects at risk, i.e. the effective sample size, for estimating the hazard is always small in the tail of the distribution.

Since it is difficult to estimate the variance of the rank-based estimator well by conventional methods, Jin et al. (2003) use a resampling method, based on Jin et al. (2001), in which one needs to generate repeatedly some nonnegative random variables and to solve the randomly perturbed optimisation problems. The variance estimator depends on the random number generator, the number of iterations and the nature of the perturbation. For large numbers of iterations it can be time-consuming to solve the optimisation problems. Also, it is not clear what is the preferred distribution for the perturbation. For example, Jin et al. (2001) have used gamma and beta random variables.

The empirical likelihood method was proposed by Thomas & Grunkemeier (1975) to obtain better confidence intervals in connection with the Kaplan-Meier estimator. Owen (1988, 1990) and many others developed this into a general methodology. It has many desirable statistical properties (Owen, 2001). Recently, the empirical likelihood method has been shown to work in various inference problems involving censored/truncated data. One nice feature of the empirical likelihood method is that we can compute p-values of a test and construct confidence intervals without estimating the variance of the statistic. The test statistic can be referred to a central chi-squared distribution under null hypothesis. It can be very difficult to estimate the variances of the statistics as in the case of the rank-based regression estimator for the censored accelerated failure time model.

We propose in this paper an empirical likelihood testing procedure for the censored rank regression estimator where the likelihood is defined as the censored empirical likelihood of the error variables and we show the limiting distribution of the log empirical likelihood ratio is a central chi-squared distribution under null hypothesis. The empirical likelihood method avoids the need to estimate the variance; instead one must carry out a constrained maximisation of the censored empirical likelihood, which can be done re-

liably. Furthermore, to test one hypothesis or obtain a p-value, the empirical likelihood method involves solving only one optimisation problem, whereas the resampling method needs to estimate the variance first by solving a very large number of repeated optimisation problems. Also, the empirical likelihood inference is 'repeatable' (same data set and hypothesis will always give an identical p-value) whereas the resampling method is not.

2. The regression model and the empirical likelihood

Consider the linear regression model

$$\log T_i = \beta_0' X_i + \epsilon_i \qquad (i = 1, \cdots, n) \ , \tag{1}$$

where the ϵ_i are independent, with an unspecified distribution $F_0(t)$. Since we are only going to use the ranks in the estimation, the mean of ϵ_i is not identifiable, and thus is not assumed to be zero. Thus, in effect, the intercept term is included in the ϵ_i. The β_0 in the above model is a q-vector of regression parameters to be estimated, and X is a matrix not including a column of 1's.

Let C_i be the censoring time for T_i. Assume C_i and T_i are independent conditionally on X_i. The data we observe consist of $(\tilde{T}_i, \delta_i, X_i)$, where

$$\tilde{T}_i = \min(T_i, C_i) \ , \qquad \delta_i = I_{[T_i \le C_i]} \ .$$

Define $e_i(b) = \log \tilde{T}_i - b' X_i$, the residuals when b is an estimator of β_0. The rank-based estimator $\hat{\beta}$ of β_0 is the solution of the following estimating equation (Jin et al., 2003):

$$0 = \sum_{i=1}^{n} \delta_i \phi\{e_i(b)\}[X_i - \bar{X}\{e_i(b)\}] \ , \tag{2}$$

where $\bar{X}\{e_i(b)\}$ is the average of those covariates, X_j, that $\log \tilde{T}_j - b' X_j \ge e_i(b)$. We assume that the function $\phi(\cdot)$ is either a constant, resulting a log-rank type estimator, or equal to the number of subjects at risk, i.e. a Gehan-type estimator, or some other predictable function.

We define a censored empirical likelihood, for the censored $e_i(b)$'s, as

$$EL = \prod_{i=1}^{n} p_i^{\delta_i} (1 - \sum_{e_j \le e_i} p_j)^{1-\delta_i} \ , \tag{3}$$

where $p_i \geq 0$ and $\sum p_i = 1$. The corresponding constraint equation derived from the rank estimator to be used with the censored empirical likelihood is

$$0 = \sum_{i=1}^{n} \phi\{e_i(b)\} \frac{[X_i - \bar{X}\{e_i(b)\}]}{n w_i} \ \delta_i p_i \ , \tag{4}$$

where w_i is the jump size of the Kaplan-Meier estimator at $e_i(b)$. The Kaplan-Meier estimator is computed from $(e_j(b), \delta_j)$, $j = 1, \cdots, n$. We denote this Kaplan-Meier estimator by $\hat{F}_{KM}(b, t)$.

The empirical likelihood ratio is obtained as follows. The denominator of the ratio can easily be obtained by letting $p_i = \Delta \hat{F}_{KM}\{\hat{\beta}, e_i(\hat{\beta})\}$ in (3). The numerator of the empirical likelihood ratio is obtained by maximising (3) with respect to p_i subject to the linear constraint (4). This constrained optimisation does not have an explicit analytical solution but can be computed reliably by, for example, a generalised EM algorithm. The E-step is the same as in Turnbull (1974) and the M-step is a weighted version of constrained maximisation similar to Owen (1990) Theorem 1. For more details, see Zhou (2004). This will give $p_i = \tilde{p}_i$, insertion of which into (3) provides the numerator.

In the Appendix we prove the following theorem.

THEOREM 1. *Under mild regularity conditions the Wilks theorem holds for the empirical likelihood ratio for testing the hypothesis $H_0 : \beta = \beta_0$ versus $H_A : \beta \neq \beta_0$; that is*

$$\lambda := -2 \log ELR(\beta_0) \longrightarrow \chi^2_q \qquad \text{as } n \to \infty$$

in distribution when the null hypothesis is true, where $ELR(\beta_0)$ denotes the empirical likelihood ratio for $b = \beta_0$.

With the empirical likelihood ratio we can easily compute a p-value by appealing to the chi-squared quantile and the above Theorem. A 95% confidence region for β_0 can

be obtained as the collection of the b values such that the corresponding test problem, $H_0 : \beta = b$, has p-values larger than 0.05.

Remark 1. When $b = \hat{\beta}$, then clearly the constrained maximisation of the censored empirical likelihood (3) is achieved by the Kaplan-Meier estimator: $p_i = \Delta \hat{F}_{KM}\{\hat{\beta}, e_i(\hat{\beta})\}$. This $\hat{\beta}$ and p_i also solves the constraint equation (4). This imply that the empirical likelihood ratio is equal to one so that the p-value for the null hypothesis that $\beta_0 = \hat{\beta}$ is one and the confidence region for β_0 is 'centred' at $\hat{\beta}$.

Remark 2. To compute the numerator of the empirical likelihood ratio, we require a distribution F or p_i such that (i) it has support only on the uncensored $e_i(\beta_0)$'s, (ii) it satisfies the estimating equation (4), and (iii) among those F we find one that maximises the censored empirical likelihood (3).

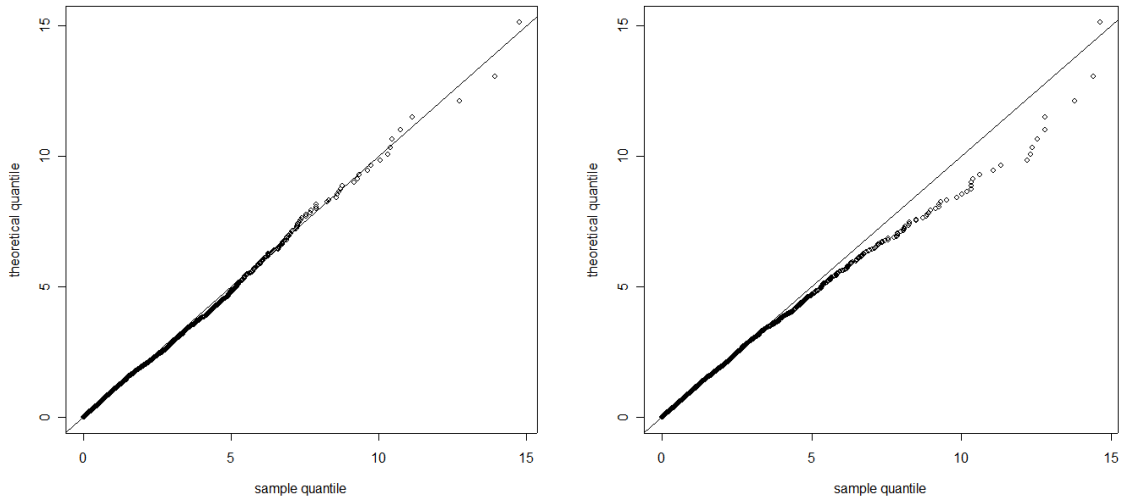
Remark 3. When maximising (3) subject to the constraint (4), we are only allowed to change the p_i's; the w_i's should always remain unchanged for a fixed b.

Remark 4. Even though the weight function in (4) is more complicated and depends on the sample, which calls for a new proof of the empirical likelihood theorem, computationally this constrained maximisation problem with respect to the p_i's is the same as the maximisation of the censored empirical likelihood with mean constraints: $\sum f(t_i)p_i = \mu$, where $f(t_i) = \phi(t_i)\{X_i - \bar{X}(e_i)\}/(nw_i)$ and $\mu = 0$. An implementation of the generalised EM algorithm for this computation is available inside the R package `emplik`, downloadable from http://cran.us.r-project.org.

Remark 5. When the dimension of β is $q > 1$, a 90% marginal confidence interval for β_1 can be obtained as the projection of a q-dimensional confidence region onto the β_1 axis. This q-dimensional confidence region should be constructed with level $\chi^2_1(0.9)$ not $\chi^2_q(0.9)$.

3. Simulation results for one-dimensional β

First we take the regression model $\log T_i = 2x_i + \epsilon_i$, where $x_i \sim \text{Un}(0.5,\ 1.5)$ and $\epsilon_i \sim \text{Un}(-0.5,\ 0.5)$. The values of $\log T_i$ are right-censored by $\log C_i$, with C_i generated according to $1 + 3.2Z_i$, where $Z_i \sim \text{Ex}(1)$. The tests we carry out are based on the censored responses, namely $\min(\log T_i, \log C_i) = \log\{\min(T_i, C_i)\}$ and δ_i, the censoring indicators. The sample size is 100, and the Q-Q plots are based on 5000 simulation runs. The value of λ is computed for each simulation run for the hypothesis $H_0 : \beta = 2$, and the resulting Q-Q plot shows a good fit to the χ^2_1 distribution. Simulations with other sample sizes produce similar Q-Q plots. The empirical likelihood ratios for the Gehan estimator has a better chi-squared approximation than those for the log-rank estimator, which tend slightly to undercover in the upper tails.



(a) (b)

Fig. 1. Simulation study for (a) Gehan estimator, (b) log-rank estimator.
Q-Q plot of $-2\log ELR$, based on 5000 simulation runs, with sample size = 100.

4. Two examples

We first illustrate the methodology with the Stanford Heart Transplant data. We

used the same data as in Miller & Halpern (1982) based on 184 cases. The data are also available as `stan` inside the R package `gss`.

`http://cran.us.r-project.org/src/contrib/Descriptions/gss.html`

We used only 152 cases as suggested by Miller & Halpern (1982) and fitted a model with only a linear term involving age, $\log_{10}(Y_i) = \beta \times \text{age} + \epsilon_i$, where Y_i is the survival time of the heart transplant patient.

The Gehan point estimate of β is -0.0253. Based on the empirical likelihood ratio, the 90% and 95% confidence intervals for β are $(-0.04177, -0.006275)$ and $(-0.04463, -0.003011)$ respectively. From the resampling method, the estimate of the variance for $\hat{\beta}$ is 0.0001137, based on 10000 resamplings and exponential perturbation. The Wald confidence interval with 90% and 95% confidence levels are then easily seen to be $(-0.04287, -0.007792)$ and $(-0.04623, -0.004434)$. The two sets of confidence intervals are similar to each other.

These results are also similar to those based on the Buckley-James estimate (Buckley & James 1979) from the same data and model: the estimate is $\hat{\beta} = -0.01990$ with 95% confidence interval $(-0.0357, -0.0028)$. See a University of Kentucky technical report by M. Zhou & G. Li for details.

For the second example we use the multiple myeloma data which is also used by Jin et al. (2003). The data can be found in SAS/STAT User's Guide (1999, pp. 2608-17, 2536-641) and are also available at

`http://ftp.sas.com/techsup/download/sample/samp_lib/statsampExamples_of_Coxs_Model.html`

There are 65 cases, of whom 17 are censored. We fitted a model with two predictors, namely the logarithm of blood urea nitrogen, log(BUN), and haemoglobin, HGB. The Gehan estimates of the regression coefficients are -15.011 and 1.318.

The resampling estimates of the variances of $(\hat{\beta}_1, \hat{\beta}_2)$ based on exponential perturbation and 10000 resamples are 36.890 and 0.814 respectively with a covariance of -1.095.

The confidence region for β_1, β_2 obtained by the empirical likelihood method is shown in Fig. 2. From the plot we can see that the correlation coefficient between the two estimators is small and the normal approximation to the joint distribution of $(\hat{\beta}_1, \hat{\beta}_2)$ is not very good.

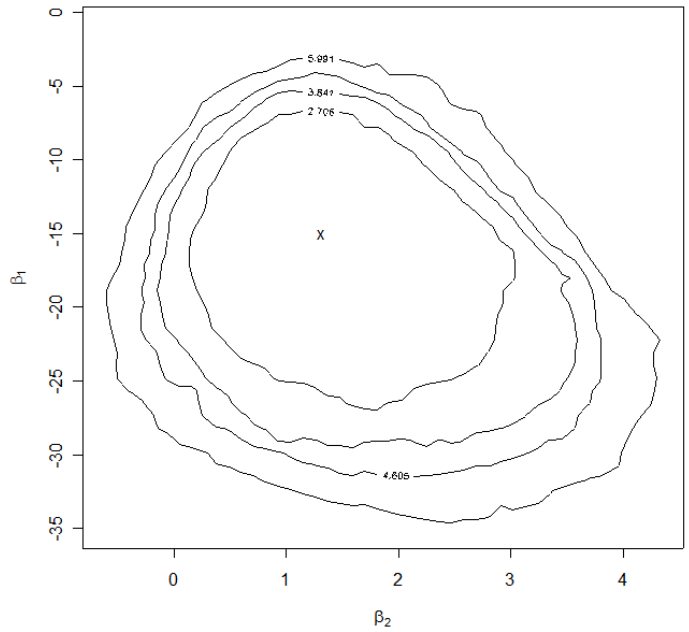


Fig. 2. Multiple myeloma data. Contour plot of $-2 \log ELR$ for (β_1, β_2).

The outer two loops are the 95% and 90% confidence regions for (β_1, β_2) jointly. The inner two loops are used to project onto the x- or y-axis to obtain marginal 95% and 90% confidence intervals for β_2 or β_1 individually. For example, the 90% confidence interval for β_2 is the x-axis shadow of the inner most loop, which has level $= \chi^2_1(0.9)$. The 95% and 90% confidence intervals for β_1 are $(-29.5, -5.3)$ and $(-27, -6.7)$, and the corresponding confidence intervals for β_2 are $(-0.15, 3.6)$ and $(0.14, 3.05)$.

Note that the contours are jagged and the enclosed regions are not convex. This is because the rank estimating function is not monotone, a fact noted by many previous authors.

Acknowledgement

I thank Z. Jin for the S-plus function which is used here to compute the Gehan rank estimator and the resampling variance estimator. I also want to thank the Editor whose careful and constructive suggestions greatly improved the presentation.

Appendix

Large sample properties of the empirical likelihood ratio

The empirical likelihood function used in (3) is exactly the same as the censored empirical likelihood based on independent, identically distributed right-censored observations used by Thomas & Grunkemeier (1975), Li (1995) and many others. Our empirical likelihood function should be considered as the likelihood of the ϵ_i, which are independent and identically distributed.

Our constraint equation, (4), however, is slightly different from the mean constraint, $\int f(t)dF(t) = \mu$, considered by Murphy & Van der Vaart (1997) and Pan & Zhou (1999): the $f(\cdot)$ we use depends on the data, and thus should be denoted by $f_n(\cdot)$, so that the constraint equation is $\int f_n(t)dF(t) = 0$. We need a generalisation of the empirical likelihood Theorem for censored data that allows a predictable integrand function, f_n, as given in Theorem 1 of a University of Kentucky technical report by M. Zhou & G. Li.

Suppose that the $\log T_i - \beta_0' X_i$ are independent with a common distribution. Based on the right censored observations $(\log \tilde{T}_i - \beta_0' X_i, \delta_i)$ we can form the Kaplan-Meier estimator $\hat{F}_{KM}(t)$ and it is well known that $\{\hat{F}_{KM}(t) - F_0(t)\}/\{1 - F_0(t)\}$ is a martingale with respect to $\mathcal{F}_t$, where $\mathcal{F}_t$ is the usual counting process filtration: see for example Fleming & Harrington (1991, p. 91).

Lemma A1. *The log-rank and Gehan weight functions used in (4), $f_n^L(t_i) = \{X_i - \bar{X}(t_i)\}/(nw_i)$ and $f_n^G(t_i) = (\sum_{j=1}^n I_{[e_j \geq t_i]})\{X_i - \bar{X}(t_i)\}/(nw_i)$, are both $\mathcal{F}_t$-predictable.*

Proof: Note that whether or not the Kaplan-Meier estimator jumps at t is not predictable but we are only concerned here with the size of the jump, if there is one. The size of the next jump of the Kaplan-Meier estimator can be computed from the history and thus is predictable. To be more specific, the jump size of the Kaplan-Meier estimator at time t, if there is one, is equal to $1/n \times 1/\{1 - \hat{G}(t-)\}$, where $\hat{G}$ is the Kaplan-Meier estimator when we reverse the censoring indicators. Therefore w_i is predictable. Clearly $\phi^L(t) = \sum_{j=1}^n I_{[e_j \geq t]}$ is predictable and $\bar{X}(t) = (\sum X_j I_{[e_j \geq t]})/(\sum I_{[e_j \geq t]})$ is also predictable. This implies that the functions f^G and f^L are predictable. $\diamond$

THEOREM 1. *When $b = \beta_0$, the residuals $\log T_i - \beta_0' X_i$ are independent and identically distributed, the e_i's are censored residuals and the estimating equation*

$$E[\phi(t)\{X - \bar{X}(t)\}] \equiv 0 \qquad (A1)$$

holds true.

Assume that the variance of the independent and identically distributed errors ϵ_i is finite and positive. Then, by the generalised empirical likelihood theorem, we have that

$$-2 \log ELR(\beta_0) \longrightarrow \chi_q^2$$

in distribution as $n \to \infty$.

Proof: In view of Lemma A1 above and the generalised empirical likelihood theorem of the technical report of M. Zhou and G. Li, we only need to verify (A1), which in turn is easily seen to be true if we first condition on $\mathcal{F}_t$ and note that $\bar{X}(t) = E(X|e_j \geq t)$ and that ϕ is predictable. $\diamond$

References

Buckley, J. & James, I. (1979). Linear regression with censored data. *Biometrika* **66** 429-36.

Fleming, T. & Harrington, D. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Gentleman, R. & Ihaka, R. (1996). R: A Language for data analysis and graphics. *J. Comp. Graph. Statist.* **5**, 299-314.

Jin, Z., Lin, D. Y., Wei, L. J. & Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341-53.

Jin, Z., Ying, Z. & Wei, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88**, 381-90.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd ed. New York: Wiley.

Lai, T. L. & Ying, Z. (1991). Rank regression methods for left-truncated and right-censored data. *Ann. Statist.* **19**, 531-56.

Li, G. (1995). On nonparametric likelihood ratio estimation of survival probabilities for censored data. *Statist. Prob. Letters* **25**, 95-104.

Miller, R. G. & Halpern, J. (1982). Regression with censored data. *Biometrika* **69**, 521-31.

Murphy, S. & van der Vaart, A. (1997). Semiparametric likelihood ratio inference. *Ann. Statist.* **25**, 1471-509.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-49.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90-120.

Owen, A. (2001). *Empirical Likelihood*. London: Chapman and Hall.

Pan, X. R. & Zhou, M. (1999). Using one parameter sub-family of distributions in empirical likelihood with censored data. *J. Statist. Plan. Infer.* **75**, 379-92.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65**, 167-79.

SAS Institute, Inc. (1999). SAS/STAT User's Guide, Version 8. Cary, NC: SAS Institute Inc.

Thomas, D. R. & Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Am. Statist. Assoc.* **70**, 865-71.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18**, 354-72.

Turnbull, B. (1974). Non-parametric estimation of a survivorship function with doubly censored data. *J. Am. Statist. Assoc.* **69**, 169-73.

Wei, L. J., Ying, Z. & Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77**, 845-51.

Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statist. Med.* **11**, 1871-9.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21**, 76-99.

Zhou, M. (2004). Empirical likelihood ratio with arbitrarily censored/truncated data by a modified EM algorithm. To appear in *J. Comp. Graph. Statist.*