The International Journal of Biostatistics

Volume 2, Issue 1

2006

Article 10

An Improved Akaike Information Criterion for Generalized Log-Gamma Regression Models

Xiaogang Su, *University of Central Florida*
Chih-Ling Tsai, *University of California at Davis*

Recommended Citation:

Su, Xiaogang and Tsai, Chih-Ling (2006) "An Improved Akaike Information Criterion for Generalized Log-Gamma Regression Models," *The International Journal of Biostatistics*: Vol. 2: Iss. 1, Article 10.

DOI: 10.2202/1557-4679.1032

An Improved Akaike Information Criterion for Generalized Log-Gamma Regression Models

Xiaogang Su and Chih-Ling Tsai

Abstract

We propose an improved Akaike information criterion (AICc) for generalized log-gamma regression models, which include the extreme-value and normal regression models as special cases. Moreover, we extend our proposed criterion to situations when the data contain censored observations. Monte Carlo results show that AICc outperforms the classical Akaike information criterion (AIC), and an empirical example is presented to illustrate its usefulness.

KEYWORDS: parametric accelerated failure time models, AICc, Kullback-Leibler information, survival model selection

Author Notes: Xiaogang Su is Assistant Professor of Statistics, Department of Statistics and Actuarial Science, University of Central Florida, Orlando, FL 32816. Chih-Ling Tsai is Professor of Statistics, Graduate School of Management, University of California, Davis, CA 95616 and Guanghua School of Management Peking University, P. R. China, 100871. The helpful and constructive comments of the Editor and the two referees are gratefully acknowledged.

1 Introduction

Over the last three decades, survival regression models have been widely used in the areas of medicine, biology, engineering, economics, and business. Two broad classes of survival regression models are in common usage: Cox (1972) proportional hazards models and accelerated failure time (AFT) models. To make a valid inference from the fitted survival model, it is important to determine the most relevant variables a priori. In Cox proportional hazards models, Klein and Moeschberger (2003, p. 277) and Collett (2003, p. 81) directly adopted the Akaike information criterion (AIC , Akaike, 1973), derived via the Kullback-Leibler distance, to find the best model. An alternative selection criterion can be found via the Bayesian approach (e.g., see Volinsky and Raftery, 2000).

In contrast to distribution-free Cox models, AFT models contain a variety of useful parametric models including normal, extreme value, log-logistic, and generalized log-gamma regression models (see Kalbfleisch and Prentice, 2002; Lawless, 1982). In parametric AFT models, Bedrick et al. (2002) applied the Kullback-Leibler distance to assess predictive influence, which motivated us to employ this distance to derive AIC for choosing the relevant variables. The AIC can be viewed as a data-based approximation for the Kullback-Leibler discrepancy function between a candidate model and the true model. It has the following general form:

$$AIC = -2 \times \log\text{-likelihood} + 2 \times \text{number of parameters}.$$

The smaller AIC results in the better candidate model. However, it is known that AIC tends to overfit, especially when the sample size is small or the number of parameters is a moderate to large fraction of the sample size. See McQuarrie and Tsai (1998) and Burnham and Anderson (2002) for detailed discussions. To amend the overfitting deficiency of AIC , Hurvich and Tsai (1989) introduced an improved information criterion, AIC_C , in Gaussian regression model selection via the Kullback-Leibler distance. As a result, this led us to extend Hurvich and Tsai's approach to obtain the AIC_C for parametric AFT models.

We organize this paper as follows. In Section 2, we develop the AIC_C for parametric AFT models, and provide a closed form of AIC_C for generalized

log-gamma regression model selections. In Section 3, we obtain an improved Akaike information criterion (AIC_C^*) for the generalized log-gamma regression model with censored observations. Section 4 presents Monte Carlo studies, which show that AIC_C and AIC_C^* outperform AIC . An empirical example is analyzed in Section 5 to illustrate the usefulness of AIC_C^* . Section 6 gives concluding remarks.

2 The AIC_C for Uncensored Data

2.1 Accelerated Failure Time Models

Suppose that the data y_i 's are generated from the true model

$$y_i = x'_{0i}\beta_0 + \sigma_0\varepsilon_i, \quad (i = 1, \dots, n) \quad (1)$$

where $y_i = \log(t_i)$, t_i is the i -th failure time, x_{0i} is a $p_0 \times 1$ vector of explanatory variables, β_0 is a $p_0 \times 1$ vector of unknown parameters, and σ_0 is a scale parameter. Furthermore, ε_i are independent identically distributed (iid) random variables with a known probability density function g . Thus, the true probability density function of y_i is

$$f_0(y_i; \beta_0, \sigma_0) = g\{(y_i - x'_{0i}\beta_0)/\sigma_0\}/\sigma_0.$$

In practice, the true model is unknown. Therefore, we fit the data with the candidate model

$$y_i = x'_i\beta + \sigma e_i, \quad (2)$$

where x_i is a $p \times 1$ vector of explanatory variables, β is a $p \times 1$ vector of unknown parameters, and σ is a scale parameter. The probability density function of e_i is the same g as defined above except that ε_i is replaced by e_i . Thus, the resulting probability density function of the candidate model is

$$f(y_i; \beta, \sigma) = g\{(y_i - x'_i\beta)/\sigma\}/\sigma.$$

To select the best model from a family of candidate models, we next derive the improved Akaike information criterion.

2.2 The AIC_C Criterion

To assess the discrepancy between the true and candidate models, we consider the Kullback-Leibler discrepancy function:

$$\Delta(\beta, \sigma) = E_0\{-2 \log f(Y; \beta, \sigma)\},$$

where $f(Y; \beta, \sigma) = \prod_{i=1}^n f(y_i; \beta, \sigma)$, $Y = (y_1, \dots, y_n)'$, and E_0 denotes the expectation evaluated under the true model, $f_0(Y; \beta_0, \sigma_0) = \prod_{i=1}^n f_0(y_i; \beta_0, \sigma_0)$. Then, we adapt the assumption used by Linhart and Zucchini (1986, p. 245) and Burnham and Anderson (2002, p. 375), namely that the candidate model family includes the true model as a special case. It should be stressed that this assumption is only made to facilitate the derivation of the criterion. Under this assumption, the columns of x_i ($i = 1, \dots, n$) can be rearranged so that $x'_{0i}\beta_0 = x'_i\beta^*$, where $\beta^* = (\beta'_0, \beta'_1)'$, and β_1 is a $(p - p_0) \times 1$ vector of zeros. In addition, the second-order Taylor expansion of the Kullback-Leibler information at $\beta = \beta^*$ is (omitting irrelevant constants):

$$\Delta(\beta, \sigma) \simeq \Delta(\beta^*, \sigma) + \dot{\Delta}(\beta^*, \sigma)(\beta - \beta^*) + (\beta - \beta^*)' \ddot{\Delta}(\beta^*, \sigma)(\beta - \beta^*)/2, \quad (3)$$

where

$$\begin{aligned} \Delta(\beta^*, \sigma) &= E_0\{-2 \log f(Y; \beta, \sigma)|_{\beta=\beta^*}\} = n \log(\sigma^2) - 2E_0\{A(\varepsilon, \sigma)\}, \\ \dot{\Delta}(\beta^*, \sigma) &= E_0[\partial\{-2 \log f(Y; \beta, \sigma)\}/\partial\beta|_{\beta=\beta^*}] = 2E_0\{Z(\varepsilon, \sigma)'X\}, \\ \ddot{\Delta}(\beta^*, \sigma) &= E_0[\partial^2\{-2 \log f(Y; \beta, \sigma)\}/\partial\beta\partial\beta'|_{\beta=\beta^*}] = 2E_0\{X'W(\varepsilon, \sigma)X\}, \end{aligned}$$

$\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$, $\varepsilon_i = (y_i - x'_i\beta^*)/\sigma_0$ for $i = 1, \dots, n$, $X = (x_1, \dots, x_n)'$, $A(\varepsilon, \sigma) = \sum_{i=1}^n \log\{g(\varepsilon_i\sigma_0/\sigma)\}$, $Z(\varepsilon, \sigma) = \frac{1}{\sigma}(\dot{g}(\varepsilon_1\sigma_0/\sigma)/g(\varepsilon_1\sigma_0/\sigma), \dots, \dot{g}(\varepsilon_n\sigma_0/\sigma)/g(\varepsilon_n\sigma_0/\sigma))'$, and $W(\varepsilon, \sigma) = \frac{1}{\sigma^2} \text{Diag}\{\{\dot{g}(\varepsilon_i\sigma_0/\sigma)/g(\varepsilon_i\sigma_0/\sigma)\}^2 - \ddot{g}(\varepsilon_i\sigma_0/\sigma)/g(\varepsilon_i\sigma_0/\sigma)\}$. For the sake of simplicity, we center the x variables so that $\dot{\Delta}(\beta^*, \sigma) = 0$.

A reasonable criterion for judging the quality of candidate models with respect to the data is $E_0\{\delta(\hat{\beta}, \hat{\sigma})\}$, where $\delta(\hat{\beta}, \hat{\sigma}) = \Delta(\beta, \sigma)|_{\beta=\hat{\beta}, \sigma=\hat{\sigma}}$, and $\hat{\beta}$ and $\hat{\sigma}$ are the maximum likelihood estimators of β and σ , respectively. Applying Equation (3), we have

$$\delta(\hat{\beta}, \hat{\sigma}) \approx \delta(\beta^*, \hat{\sigma}) + (\hat{\beta} - \beta^*)' \ddot{\Delta}(\beta^*, \hat{\sigma})(\hat{\beta} - \beta^*)/2, \quad (4)$$

where $\delta(\beta^*, \hat{\sigma}) = \Delta(\beta^*, \sigma)|_{\sigma=\hat{\sigma}} = n \log \hat{\sigma}^2 + C_1|_{\sigma=\hat{\sigma}}$, $C_1 = -2E_0\{A(\varepsilon, \sigma)\}$, and $\ddot{\delta}(\beta^*, \hat{\sigma}) = \ddot{\Delta}(\beta^*, \sigma)|_{\sigma=\hat{\sigma}}$. Replacing $\hat{\beta} - \beta^*$ in (4) by its first order approximation, $B = \{\ddot{\Delta}(\beta^*, \sigma_0)/2\}^{-1} [\{\partial\{\log f(Y; \beta, \sigma)\}/\partial\beta\}|_{\beta=\beta^*, \sigma=\sigma_0} = [E_0\{X'W(\varepsilon, \sigma_0)X\}]^{-1}X'Z(\varepsilon, \sigma_0)]$, we then can approximate the second term of (4) by $C_2|_{\sigma=\hat{\sigma}} = B'[E_0\{X'W(\varepsilon, \sigma)X\}]|_{\sigma=\hat{\sigma}}B$. This leads to

$$E_0\{\delta(\hat{\beta}, \hat{\sigma})\} \approx E_0\{n \log \hat{\sigma}^2 + (C_1 + C_2)|_{\sigma=\hat{\sigma}}\}.$$

Since $E_0[\partial^2\{-2 \log f(Y; \beta, \sigma)\}/\partial\beta\partial\sigma|_{\beta=\beta^*, \sigma=\sigma^*}] = 0$, $\hat{\beta}$ and $\hat{\sigma}$ are asymptotically independent (see Lindsey, 1996, p. 202). Hence $E_0(C_2|_{\sigma=\hat{\sigma}})$ can be approximated by $E_0(C_2^*)$, where $C_2^* = \text{tr}\{[E_0\{X'W(\varepsilon, \sigma)X\}]|_{\sigma=\hat{\sigma}}E_0(BB')\}$. Consequently, we obtain the improved Akaike information criterion AIC_C (omitting irrelevant constants) given below which is an approximate unbiased estimator of $E_0\{\delta(\hat{\beta}, \hat{\sigma})\}$.

$$AIC_C = n \log(\hat{\sigma}^2) + E_0(C_1|_{\sigma=\hat{\sigma}} + C_2^*). \quad (5)$$

The AIC_C is applicable for general AFT models. However, the computations of $E_0\{A(\varepsilon, \hat{\sigma})\}$ in $C_1|_{\sigma=\hat{\sigma}}$ and $E_0(BB')$ in C_2^* cannot be simplified under the true model (1). Hence, the penalty function in (5) does not have a simple form. To facilitate the application of AIC_C in model selections, we next focus on the generalized log-gamma models that are often used in practice (see Lawless, 1980), which results in AIC_C with a closed form.

2.3 The AIC_C For Generalized Log-Gamma Regression Models

In generalized log-gamma regression models, the probability density function of ε_i in Equation (1) is

$$g(\varepsilon_i; k) = \frac{k^{k-1/2}}{\Gamma(k)} \exp\{k^{1/2}\varepsilon_i - k \exp(\varepsilon_i k^{-1/2})\} \quad (6)$$

when $0 < k < \infty$, and it is

$$\phi(\varepsilon_i) = (2\pi)^{-1/2} \exp(-\varepsilon_i^2/2)$$

when $k = \infty$. In addition, the probability density function of e_i in (2) is the same as above except that ε_i is replaced by e_i .

We first discuss the case where $0 < k < \infty$. After algebraic simplification, we have $C_1 = -2n[(k - 1/2) \log(k) - \log \Gamma(k) + k\{\psi(k) - \log(k)\}\sigma_0/\sigma - \{\Gamma(\frac{\sigma_0}{\sigma} + k)/\Gamma(k)\}k^{(1-\sigma_0/\sigma)}]$, where ψ is a digamma function. Also, $W(\varepsilon, \sigma) = \exp\{k^{-1/2}(\varepsilon\sigma_0/\sigma)\}/\sigma^2$, $E_0\{W(\varepsilon, \sigma)\} = \Gamma(\frac{\sigma_0}{\sigma} + k)/\{\Gamma(k)\sigma^2\}k^{-\sigma_0/\sigma}I_{n \times n}$, $E_0\{W(\varepsilon, \sigma_0)\} = \frac{1}{\sigma_0^2}I_{n \times n}$, $Z(\varepsilon, \sigma_0)'X = -k^{1/2}(\exp(\varepsilon_1 k^{-1/2}), \dots, \exp(\varepsilon_n k^{-1/2}))'X/\sigma_0$, $E_0(BB') = (X'X)^{-1}X'(I_{n \times n}k^{-1} + J)X(X'X)^{-1}\sigma_0^2$, and J is an $n \times n$ matrix with elements 1. Applying Equation (5), we obtain the following selection criterion for the generalized log-gamma regression model:

$$AIC_C = n \log \hat{\sigma}^2 - 2n\{(k - 1/2) \log(k) - \log \Gamma(k)\} + 2nkE_0[\{\log(k) - \psi(k)\}\sigma_0/\hat{\sigma} + \Gamma(\frac{\sigma_0}{\hat{\sigma}} + k)/\Gamma(k)k^{-\sigma_0/\hat{\sigma}}\{1 + (p/2nk)\sigma_0^2/\hat{\sigma}^2\}]. \quad (7)$$

Next we consider the case where $k = \infty$. After algebraic simplification, we have $C_1 = n\sigma_0^2/\sigma^2$, and $C_2^* = p\sigma_0^2/\hat{\sigma}^2$. These results, together with Equation (5), yield the selection criterion

$$AIC_C = n \log \hat{\sigma}^2 + n(n + p)/(n - p - 2). \quad (8)$$

To compute the AIC_C given by Equation (7), we adopt Kotz and Nadarajah's (2000) approximation approach. First, applying Young and Bakir's (1987) Equation (3.7), we obtain $T = E_0(\sigma_0/\hat{\sigma}) \approx \{E_0(\hat{\sigma}/\sigma_0)\}^{-1} \approx \{(2a_1)^{1/2}[(n - p)/n]^{1/2} - 1 + (0.5p + 0.75)/n\} + 1 - (a_1p + a_2)/n\}^{-1}$, where a_1 and a_2 are functions of k and are given in Young and Bakir (1987, Table 1). Next, using the second-order Taylor expansions of $\Gamma(\sigma_0/\hat{\sigma} + k)$ and $k^{-\sigma_0^2/\hat{\sigma}^2}$ around $\sigma_0/\hat{\sigma} = 0$ and $\sigma_0^2/\hat{\sigma}^2 = 0$, respectively, the third term of Equation (7) can be simplified to $2nk + (n + p + n/4k)T^2$. Hence, AIC_C in (7) becomes

$$AIC_C = n \log \hat{\sigma}^2 - 2n\{(k - 1/2) \log(k) - \log \Gamma(k) - k\} + (n + p + n/4k)T^2. \quad (9)$$

As for the case where $k = \infty$, we have $a_1 = 0.5$, $a_2 = 0.75$, and $T^2 = n/(n - p)$, which is close to the exact result $n/(n - p - 2)$. Furthermore, using the fact that $\{(k - 1/2) \log(k) - \log \Gamma(k) - k\} \rightarrow \log(2\pi)/2$ and $n/4k \rightarrow 0$ as $k \rightarrow \infty$, we have $AIC_C = n \log \hat{\sigma}^2 + n(n + p)/(n - p)$, which is virtually identical to AIC_C as given in (8).

Remark 1. As an alternative to the approximation approach given by Equation (9), we can adopt Hurvich and Tsai's (1990) Monte Carlo approach for computing the last term of AIC_C in Equation (7) when $0 < k < \infty$. This is because $\sigma_0/\hat{\sigma}$ is a pivotal quantity and therefore its distribution is independent of the parameter β_0 and of σ_0 . This approach is also mentioned by Kotz and Nadarajah (2000, p. 41) in constructing tolerance limits for extreme value distributions. Compared to the explicit form given in Equation (9), the involvement of simulations makes this Monte Carlo approach less appealing computationally. Accordingly, we will use Equation (9) to compute AIC_C in the rest of the paper. It is worth noting that AIC_C given in (8) is the same as the improved information criterion proposed by Hurvich and Tsai (1989) when $k = \infty$.

Remark 2. When $k = \infty$, AIC_C in Equation (8) is the same as $AIC = n \log \hat{\sigma}^2 + 2p$ obtained from the definition of AIC in Section 1 plus a quantity $\delta_n(p) = (n^2 - np + 2p^2 + 4p)/(n - p - 2)$. Because AIC is asymptotically equivalent to Shibata's (1981) selection criterion, $(n+2p)\hat{\sigma}^2$, and $\delta_n(p)$ satisfies Shibata's two conditions in his Section 5, we can follow Shibata's arguments to show that AIC_C in Equation (8) is an efficient criterion (i.e., It selects the best finite dimensional candidate model in large samples when the true model is of infinite dimension). When $k < \infty$, we expand $\exp(\varepsilon_i k^{-1/2})$ in Equation (6) and then re-express the generalized log-gamma density function as

$$g(\varepsilon_i; k) = \frac{k^{k-1/2} \exp(-k)}{\Gamma(k)} \exp\{-\varepsilon_i^2/2 - \varepsilon_i^3/(6k^{1/2}) - \dots\}.$$

It can be shown that $g(\varepsilon_i; k) \rightarrow \exp(-\varepsilon_i^2/2)/(2\pi)^{1/2}$ as $k \rightarrow \infty$. Hence, AIC_C in Equation (7) is an approximately efficient criterion as k gets large. For the small k , we conjecture that AIC_C is also an efficient criterion, which needs further study.

In practice, the data may be censored, which leads us to the study of variable selection for censored observations.

3 The AIC_C for Censored Data

When the data contain right censored observations, we can apply Equation (3) to obtain the selection criterion AIC_C for AFT models. However, the resulting criterion is too general to be of practical use, and therefore we will focus only on generalized log-gamma regression models.

In a random sample of n individuals, we assume that r lifetimes and $n - r$ censoring times are observed. In addition, let y_i denote either the log-lifetimes or the log-censoring times for the i th individual, with D and C denoting the sets of individuals for whom lifetimes and censoring times, respectively, are observed. Hence, $|D| = r$ and $|C| = n - r$. For the given sample, the log likelihood function of the true model is

$$L_1(Y; \beta^*, \sigma_0) = -r \log \sigma_0 + \sum_{i \in D} \log g(\varepsilon_i; k) + \sum_{i \in C} \log Q_1(k, k e^{\varepsilon_i k^{-1/2}}), \quad (10)$$

where $0 < k < \infty$, the ε_i are as given in (1), $g(\varepsilon; k)$ is the probability density function of the generalized log-gamma defined in (6), and $Q_1(k, a) = \int_a^\infty u^{k-1} e^{-u} / \Gamma(k) du$. When $k = \infty$, the log likelihood function becomes

$$L_2(Y; \beta^*, \sigma_0) = -r \log \sigma_0 - 1/(2\sigma_0^2) \sum_{i \in D} (y_i - x'_i \beta^*)^2 + \sum_{i \in C} \log Q_2\{(y_i - x'_i \beta^*)/\sigma_0\}, \quad (11)$$

where $Q_2(a) = \int_a^\infty \phi(u) du$, and $\phi(u)$ is the standard normal density function. Usually the true model is unknown and the data are fit with the candidate model. The log likelihood functions of the candidate models, $L_1(Y; \beta, \sigma)$ for $0 < k < \infty$ and $L_2(Y; \beta, \sigma)$ for $k = \infty$, have the same form as given in equations (10) or (11) except that β^* , σ_0 , ε_i , and ε are replaced by β , σ , $e_i = (y_i - x'_i \beta)/\sigma$, and $e = (e_1, \dots, e_n)'$, respectively. We then derive the selection criterion for the case of $k = \infty$.

3.1 The AIC_C for normal distribution with censored data

Applying the linear approximation

$$Q_2(e_i) \approx \phi(e_i) \{b_1(1 + b_2 e_i)^{-1}\} \quad (12)$$

given in Abramowitz and Stegun (1970, Eq. (26.2.16)), and approximating $\log(1 + b_2 e_i)$ with $b_2 e_i$, we have

$$\log Q_2(e_i) \approx \log(b_1) - \log \sigma - e_i^2/2 - b_2 e_i,$$

where $b_1 = 0.43618$ and $b_2 = 0.33267$. After algebraic simplification, we obtain

$$\delta(\beta^*, \hat{\sigma}) \approx n \log \hat{\sigma}^2 + n \sigma_0^2 / \hat{\sigma}^2, \quad (13)$$

where $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2 .

Applying (12) and replacing its b_1 by the reasonable approximation b_2 , we have

$\partial^2 \{L_2(Y; \beta, \sigma)\} / \partial \beta \partial \beta' \approx -X'X / \sigma^2$, and thus $\ddot{\delta}(\beta^*, \hat{\sigma}) \approx 2(X'X) / \hat{\sigma}^2$. Under the true model, $\hat{\beta} - \beta^*$ is approximately multivariate normal, $N(0, \sigma_0^2 (X'X)^{-1})$, and the quantity $n^* \hat{\sigma}^2 / \sigma_0^2$ is approximately distributed as $\chi_{n^*-p}^2$ independently of $\hat{\beta}$, where $n^* = r + \sum_{i \in C} \lambda(\hat{e}_i)$, $\lambda(\hat{e}_i) = V(\hat{e}_i) \{V(\hat{e}_i) - \hat{e}_i\}$, $V(\hat{e}_i) = \phi(\hat{e}_i) / Q(\hat{e}_i)$ and $\hat{e}_i = (y_i - x_i' \hat{\beta}) / \hat{\sigma}$ (see Lawless, 1982, p. 318). Thus, $E_0(n \sigma_0^2 / \hat{\sigma}^2) \approx nn^* / (n^* - p - 2)$, and $E_0\{(\hat{\beta} - \beta^*)' \ddot{\delta}(\beta^*, \hat{\sigma}) (\hat{\beta} - \beta^*) / 2\} \approx E_0[\{(\hat{\beta} - \beta^*)' X' X (\hat{\beta} - \beta^*) / \sigma_0^2\} / (\hat{\sigma}^2 / \sigma_0^2)] = n^* p / (n^* - p - 2)$. Using these results in conjunction with equations (4), (5), and (13), we have $E_0\{\delta(\hat{\beta}, \hat{\sigma})\} \approx E_0(n \log \hat{\sigma}^2) + nn^* / (n^* - p - 2) + n^* p / (n^* - p - 2)$. Consequently, we obtain an improved model selection criterion

$$AIC_C^* = n \log \hat{\sigma}^2 + n^* (n + p) / (n^* - p - 2). \quad (14)$$

If the data have no censoring observations, then $n^* = n$ and $AIC_C^* = AIC_C$ given in Equation (8). Next, we study the selection criterion for the generalized log-gamma distribution when the data contain right censored observations.

3.2 The AIC_C for generalized log-gamma distribution with censored data

When the data are censored, Lawless (1982, p. 332) suggested using the procedure for censored normal data to obtain least squares estimators for generalized log-gamma models. Here, we will adapt this suggestion to derive a model selection criterion. We first approximate the term $\exp(\varepsilon_i k^{-1/2})$ for the generalized log-gamma distribution by its second order Taylor expansion,

$1 + \varepsilon_i k^{-1/2} + (\varepsilon_i^2 k^{-1})/2$. The resulting approximate generalized log-gamma distribution of ε_i is

$$g(\varepsilon_i; k) = \frac{k^{k-1/2}}{\Gamma(k)} \exp\{k^{1/2}\varepsilon_i - k \exp(\varepsilon_i k^{-1/2})\} \approx \frac{k^{k-1/2} \exp(-k)}{\Gamma(k)} \sqrt{2\pi} \phi(\varepsilon_i).$$

Therefore, the log likelihood function of (10) can be approximated by $\tilde{L}_2(Y; \beta^*, \sigma_0) = r\{(k - 1/2) \log(k) - \log \Gamma(k) - k\} + L_2(Y; \beta^*, \sigma_0)$, where $L_2(Y; \beta^*, \sigma_0)$ is given in (11).

Through the same approach used for obtaining Equation (14), and in conjunction with Equation (9), we obtain the selection criterion

$$AIC_C^* = n \log \hat{\sigma}^2 - 2r\{(k - 1/2) \log(k) - \log \Gamma(k) - k\} + \frac{n^*(n+p)}{(n^* - p - 2)}, \quad (15)$$

where $\hat{\sigma}^2$ is the estimator computed from the generalized log-gamma distribution with censored data. Note that AIC_C^* and AIC_C given in Equation (9) have the same first two terms, and these two criteria are approximately the same if the data have no censored observations and k becomes large.

Remark 3. In the non-censored regression model with $n > 4$ and $n > p$, we can show that the penalty function of AIC_C in (9), $(n + p + n/4k)T^2$, is larger than the penalty function of AIC , $2p$, obtained from the generic definition of AIC in Section 1. In the censored data with $n > p$, we can easily see that the penalty function of AIC_C^* in Equations (14) and (15), $n^*(n+p)/(n^* - p - 2)$, is larger than the penalty function of AIC , $2p$. As a result, AIC_C and AIC_C^* prevent more overfitting than AIC , which yield greater correct models being selected (see also simulation studies in the next section).

4 Simulation

In this section, we compare the performance of AIC_C given by Equations (8) & (9) and AIC_C^* given by Equations (14) & (15) versus the classical Akaike information criterion, $AIC = -2 \log f(Y; \hat{\beta}, \hat{\sigma}) + 2p$. Data were generated from the true model (1) with $\beta_0 = (1, 1, 2, 3)'$ and $\sigma_0 = 0.5$ and 3, respectively. The ε_i are iid generalized log-gamma random variables with density function (6).

Figure 1. Plots of penalty functions of AIC, AIC_c, and AIC_c^{*} when $n = 25$ and $\sigma = 0.5$.

Table 1

Proportion of model order selected by the minimums of AIC and AIC_C in 1000 realizations: the numbers for the ‘Selected Model Order’ denote the number of variables being selected in a nested fashion (e.g., 2 for (X₁, X₂), 3 for (X₁, X₂, X₃), and so on.)

Signal	Sample Size	AFT Model	Selection Criterion	Selected Model Order						
				1	2	3	4	5	6	7
$\sigma = 3$	$n = 15$	$k = 1$	<i>AIC</i>	1.9	2.8	36.9	12.1	10.2	12.9	23.3
			<i>AIC_C</i>	12.9	10.4	66.6	7.8	1.3	0.9	0.2
		$k = 5$	<i>AIC</i>	0.0	0.0	46.7	12.1	10.3	11.7	19.2
			<i>AIC_C</i>	0.0	0.1	85.9	8.6	3.9	1.1	0.4
		$k = \infty$	<i>AIC</i>	1.0	2.2	42.5	12.5	10.7	11.6	19.6
			<i>AIC_C</i>	11.1	7.9	73.3	6.4	0.8	0.5	0.0
	$n = 25$	$k = 1$	<i>AIC</i>	0.2	0.9	56.5	13.7	8.5	9.1	11.1
			<i>AIC_C</i>	1.3	2.1	80.9	9.9	3.5	2.00	0.4
		$k = 5$	<i>AIC</i>	0.0	0.0	60.3	13.7	8.2	7.6	10.2
			<i>AIC_C</i>	0.0	0.0	81.8	10.7	4.4	1.8	1.3
		$k = \infty$	<i>AIC</i>	0.1	0.3	59.4	14.8	8.7	7.5	9.4
			<i>AIC_C</i>	0.6	0.7	85.3	8.7	2.7	1.3	0.7
$\sigma = 0.5$	$n = 15$	$k = 1$	<i>AIC</i>	0.0	0.0	44.2	10.4	13.9	13.8	17.7
			<i>AIC_C</i>	0.0	0.0	89.3	7.7	2.5	0.4	0.1
		$k = 5$	<i>AIC</i>	0.0	0.0	47.5	12.7	8.3	11.7	19.9
			<i>AIC_C</i>	0.0	0.0	86.7	7.6	3.0	1.9	0.8
		$k = \infty$	<i>AIC</i>	0.0	0.0	45.3	13.1	10.4	12.1	19.2
			<i>AIC_C</i>	0.0	0.0	92.8	5.8	1.4	0.0	0.0
	$n = 25$	$k = 1$	<i>AIC</i>	0.0	0.0	58.5	13.8	8.2	8.9	10.6
			<i>AIC_C</i>	0.0	0.0	84.5	9.6	3.5	1.4	1.1
		$k = 5$	<i>AIC</i>	0.0	0.0	61.7	12.9	7.8	8.1	9.4
			<i>AIC_C</i>	0.0	0.0	83.1	9.6	3.8	2.4	1.1
		$k = \infty$	<i>AIC</i>	0.0	0.0	59.2	13.9	8.8	8.5	9.6
			<i>AIC_C</i>	0.0	0.0	86.0	8.4	3.5	1.4	0.7

For the sake of comparison, we consider $k = 1$ (extreme value distribution), $k = 5$ and $k = \infty$ (normal distribution).

Seven candidate variables were stored in an $n \times 7$ matrix $\tilde{X}$ of independent identically distributed standard normal random variables, $N(0,1)$. The candidate models are linear and include the columns of $X = (\mathbf{1}, \tilde{X})$ in a sequentially nested fashion, and the true model consists of the columns $\mathbf{1}$ and the first three columns of X . In addition, three censoring rates were considered: 0% (no censoring), 25%, and 50%. Two sample sizes of $n = 15$ and 25 were used for the uncensored case, and $n = 25$ and 40 for the censored case. Moreover, there were 1000 replications per simulation experiment.

Table 1 presents the proportion of the order selected by AIC and AIC_C for generalized log-gamma regression models ($k = 1, 5, \infty$) with uncensored observations. It shows that AIC_C strongly outperforms AIC across all three AFT regression models when the sample size is 15 and $\sigma_0 = 3$. We obtain a similar conclusion when the σ_0 decreases from 3 to 0.5. It is not surprising that the performances of both AIC and AIC_C improve since the true model is easily identified. For $n = 25$, the correction factor of the penalty function of AIC_C , p/n , is smaller than that for $n = 15$. Hence, the superiority of AIC_C over AIC decreases slightly, but AIC_C still outperforms AIC . As indicated by one anonymous referee, it is worthy of note that in two cases of $\sigma = 3$ and $n = 15$ when $k = 1$ and $k = \infty$, AIC_C winds up with more underfitted final selections (12.9% and 11.1% respectively). Since underfitting generally causes more concern than overfitting, we suggest that one be cautious when applying AIC_C to very small samples that involve weak signals.

In censored model selections, we consider the cases of 25% and 50% censoring data. Because the results show a similar pattern, we only present the 50% case. Table 2 gives the proportion of the order selected by AIC and AIC_C^* . It clearly indicates that AIC_C^* is superior to AIC across the three AFT models, two sample sizes, and two variances of noises. Furthermore, Figure 1 depicts the penalty functions of AIC , AIC_C and AIC_C^* when $n = 25$ and $\sigma = 0.5$. It shows that AIC_C^* has the largest penalty, whereas AIC has the smallest penalty. Consequently, AIC performs the worst as it tends to overfitting. In contrast, AIC_C^* performs the best and is also slightly better than AIC_C for uncensored model selections. This is because the penalty function of AIC_C^* is larger than that of AIC_C , which prevents more overfitting.

Remark 4. It is known that overfitting inflates the variance of estimations. In contrast, underfitting yields the bias in the parameter estimators. Both overfitting and underfitting are undesirable for data analysis. A criterion that can balance the tendencies to overfit and underfit is preferable. To this end, we proposed the selection criteria AIC_C and AIC_C^* that balance between complexity and goodness-of-fit. Tables 1 and 2 show that when $\sigma = 3$ (i.e., the signal is weak) and $n = 15$, AIC_C and AIC_C^* tend to underfit more often than AIC . This finding is not surprising since the true model is weakly identifiable and the sample size is small. In this case, however, AIC_C and AIC_C^* identify much more correct models than those of AIC . Moreover, the percentage of underfitting becomes negligible (or zero) as the sample size increases to $n = 25$ (or σ decreases to 0.5). Detailed discussions on underfitting and overfitting can be found in McQuarrie and Tsai (1998), Burnham and Anderson (2004), and Seber and Lee (2003).

Since AIC_C was derived with fixed k , we also conducted simulations to investigate its performance with a mis-specified k . For example, with models generated from the extreme value distribution $k = 1$, we also fit generalized log-gamma models with $k = 5$ and $k = \infty$. The results are not reported here. We found that AIC , AIC_C , and AIC_C^* all show considerable robustness with respect to mis-specification of k . Furthermore, both AIC_C and AIC_C^* outperform AIC consistently in all the model configurations considered above.

5 Example: Ovarian Cancer Study

Following surgical treatment of ovarian cancer, Edmunson et al. (1979) studied the anti-tumor effects of two different forms of chemotherapy treatment, cyclophosphamide alone and cyclophosphamide combined with adriamycin. The data set was obtained from Therneau (1986), and consists of 26 women with minimal residual disease who had experienced surgical excision. After surgery, the patients were further classified according to whether the residual disease was completely or partially excised. The response variable is $Y = \log(T)$, where T is the survival time in days. The explanatory variables are X_1 (1= single treatment, 2=combined treatment), X_2 (Age), X_3 (1=incomplete residual disease, 2=complete residual disease), and X_4 (1=good performance, 2=poor

Table 2

Proportion of model order selected by the minimums of AIC and AIC_C^ in 1000 realizations with 50% censoring data: the numbers for the ‘Selected Model Order’ denote the number of variables being selected in a nested fashion (e.g., 2 for (X_1, X_2) , 3 for (X_1, X_2, X_3) , and so on.)*

Signal	Sample Size	AFT Model	Selection Criterion	Selected Model Order						
				1	2	3	4	5	6	7
$\sigma = 3$	$n = 25$	$k = 1$	AIC	4.8	3.3	43.8	12.9	10.4	10.3	14.4
			AIC_C^*	6.9	5.0	79.5	5.0	1.7	0.3	1.5
		$k = 5$	AIC	0.0	0.1	55.9	15.9	8.3	7.2	12.5
			AIC_C^*	0.0	0.0	90.5	6.6	0.7	0.9	1.3
		$k = \infty$	AIC	0.9	1.9	51.9	12.4	7.8	9.0	16.1
			AIC_C^*	1.0	2.0	88.5	6.1	1.3	0.6	0.5
	$n = 40$	$k = 1$	AIC	0.6	0.8	59.7	13.8	8.6	8.3	8.3
			AIC_C^*	0.1	0.4	81.5	11.4	4.6	1.5	0.5
		$k = 5$	AIC	0.0	0.0	63.8	13.7	8.6	7.5	6.4
			AIC_C^*	0.0	0.0	82.0	10.2	4.2	2.6	1.0
		$k = \infty$	AIC	0.0	0.5	66.6	12.6	7.1	6.9	6.4
			AIC_C^*	0.0	0.2	82.2	10.8	4.1	1.1	1.6
$\sigma = 0.5$	$n = 25$	$k = 1$	AIC	0.0	0.0	51.3	12.4	10.5	10.8	15.1
			AIC_C^*	0.0	0.0	88.7	5.8	1.8	1.4	2.4
		$k = 5$	AIC	0.0	0.0	54.1	12.8	11.8	7.7	13.7
			AIC_C^*	0.0	0.0	90.3	5.0	2.5	1.0	1.2
		$k = \infty$	AIC	0.0	0.0	55.8	11.9	9.2	8.8	14.3
			AIC_C^*	0.0	0.0	89.4	6.9	1.9	0.4	1.4
	$n = 40$	$k = 1$	AIC	0.0	0.0	62.1	13.4	7.2	8.5	8.9
			AIC_C^*	0.0	0.0	81.8	10.3	3.7	2.8	1.5
		$k = 5$	AIC	0.0	0.0	64.0	14.3	6.5	6.8	8.3
			AIC_C^*	0.0	0.0	83.8	10.5	3.4	1.6	0.7
		$k = \infty$	AIC	0.0	0.0	63.7	13.2	8.8	6.0	8.3
			AIC_C^*	0.0	0.0	81.4	9.6	5.2	2.7	1.2

Table 3*The best models selected by minimum AIC and AIC_C^{*} for ovarian cancer data*

variables	$k = 1$		$k = 5$		$k = \infty$	
	<i>AIC</i>	<i>AIC_C[*]</i>	<i>AIC</i>	<i>AIC_C[*]</i>	<i>AIC</i>	<i>AIC_C[*]</i>
intercept only	61.535	48.793	60.443	59.967	59.870	65.944
X_1	62.355	51.327	60.793	61.053	59.708	65.259
X_2	47.566	33.221	46.918	41.298	46.945	46.006
X_3	59.382	49.216	57.761	57.892	56.799	62.005
X_4	62.849	52.069	62.119	63.382	61.773	69.435
X_1, X_2	47.126	32.316	45.653	38.971	44.812	41.337
X_1, X_3	59.692	51.539	57.951	58.741	56.975	62.080
X_1, X_4	63.796	55.383	62.449	65.067	61.521	69.295
X_2, X_3	47.663	34.486	46.504	42.057	45.841	45.154
X_2, X_4	49.518	37.275	48.826	44.785	48.684	48.533
X_3, X_4	60.487	53.579	58.991	62.095	58.189	66.296
X_1, X_2, X_3	47.406	34.877	45.496	40.638	44.285	42.058
X_1, X_2, X_4	49.109	37.512	47.607	43.362	46.710	45.100
X_1, X_3, X_4	61.085	57.383	59.146	63.825	58.162	66.896
X_2, X_3, X_4	49.637	40.436	48.494	47.277	47.841	49.779
X_1, X_2, X_3, X_4	49.333	41.749	47.462	46.827	46.272	47.748

performance). The censored indicator variable is X_0 (0= censored observation, 1=uncensored), and there are a total of 12 observations that were censored.

We select the best model from all possible $2^4 - 1 = 15$ candidate models. For $k = 1, 5$ and ∞ , Table 3 presents best models selected via *AIC* and *AIC_C^{*}*. For $k = 1$, both *AIC* and *AIC_C^{*}* select variables X_1 and X_2 . When $k = 5$ and $k = \infty$, *AIC_C^{*}* selects variables X_1 and X_2 , whereas *AIC* chooses variables X_1, X_2 , and X_3 . In this case, *AIC* selects one more variable than *AIC_C^{*}*, a tendency towards overfitting that is consistent with our simulation findings.

To assess the performance of all five best models selected via *AIC* and *AIC_C^{*}* across $k = 1, 5$ and ∞ , Table 4 presents parameter estimates, the

standard errors of parameter estimates, and the χ^2 Wald test statistics with their corresponding p-values. In the extreme value model fitting, the p-values of Treatment (X_1) and Age (X_2) indicate that they play slightly more important roles than those resulting from the Weibull model fitting (see Collet, 2003, p. 188). However, the Treatment is only significant at the 10% level. When fitting the generalized log-gamma model with $k = 5$, X_3 is apparently redundant, and both X_1 and X_2 are of higher significance than those in the extreme value model fitting. As k increases to ∞ , X_3 is still redundant, while both X_1 and X_2 have p-values of the highest significance among all five model fittings. In particular, the treatment effect is significant at the 5% level. To determine the best choice of k among the three models selected by AIC_C^* , we computed their corresponding log-likelihood scores, -20.563 , -19.826 , and -19.406 . The normal regression model with $k = \infty$ has the largest likelihood score of -19.406 .

In conclusion, the normal model with variables X_1 and X_2 selected by AIC_C^* should be used to assess the magnitude of the treatment effect as well as the impact of age on the log-scaled survival times of ovarian cancer patients.

6 Discussion

We propose an improved Akaike information criterion to select regression variables in generalized log-gamma models, which represent an important family of parametric accelerated failure time (AFT) models. It enables the prevention of overfitting problems encountered by the classical Akaike information criterion when the sample size is small or when the number of fitted parameters is a moderate to large fraction of the sample size. Simulation studies show that both AIC_C and AIC_C^* outperform AIC in a number of model configurations.

It is worth noting that AIC_C and AIC_C^* are only applicable to selecting regression variables for the given index k . In unreported simulations, however, we found that AIC_C shows considerable robustness even with a misspecified k . Because this finding lacks of theoretical justifications, it would be interesting to extend the improved AIC criterion to jointly select the index k and regression variables. In addition, Robins and Tsiatis (1992) and Robins (1992) have considered a class of semiparametric AFT models by allowing the distribution

Table 4

Results of the five best fitting models selected by minimum AIC and AIC_C^ , respectively, to the ovarian cancer data.*

Extreme Value Model: $k = 1$				
$\log(T) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sigma \varepsilon$				
	estimate	s.e.	χ^2	p-value
β_0	10.425	1.434	52.86	< .0001
β_1	0.562	0.340	2.73	0.099
β_2	-0.079	0.020	15.97	< 0.0001
log-likelihood score:				-20.563
Generalized Log-Gamma Model: $k = 5$				
$\log(T) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \sigma \varepsilon$				
	estimate	s.e.	χ^2	p-value
β_0	10.422	1.237	71.02	< .0001
β_1	0.564	0.309	3.34	0.0677
β_2	-0.068	0.017	14.95	0.0001
β_3	-0.515	0.352	2.14	0.1432
log-likelihood score:				-18.748
$\log(T) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sigma \varepsilon$				
	estimate	s.e.	χ^2	p-value
β_0	10.162	1.264	64.60	< 0.0001
β_1	0.620	0.327	3.59	0.0581
β_2	-0.079	0.018	18.31	< 0.0001
log-likelihood score:				-19.826
Normal Model: $k = \infty$				
$\log(T) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \sigma \varepsilon$				
	estimate	s.e.	χ^2	p-value
β_0	10.293	1.141	81.41	< 0.0001
β_1	0.598	0.296	4.08	0.0433
β_2	-0.068	0.016	17.97	< 0.0001
β_3	-0.525	0.326	2.59	0.1076
log-likelihood score:				-18.142
$\log(T) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sigma \varepsilon$				
	estimate	s.e.	χ^2	p-value
β_0	9.839	1.128	76.07	< 0.0001
β_1	0.684	0.315	4.73	0.0297
β_2	-0.077	0.017	20.36	< 0.0001
log-likelihood score:				-19.406

of the error term unspecified. Since the semiparametric AFT model involves less statistical assumptions, they have recently become attractive for statistical researchers. For example, Jin et al. (2003) investigated rank-based estimation of semi-parametric AFT models, while Walker and Mallick (1999), Hanson and Johnson (2004), and Ghosal and Ghosh (2005) studied Bayesian approaches. Thus, developing variable selection criteria for semiparametric AFT models would be another interesting topic for future research.

REFERENCES

- Abramowitz, M., and Stegun, I. A. (1970). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csaki), Budapest: Akademiai Kiado, pp. 267-281.
- Bedrick, E. J., Exuzides, A., Johnson, W. O., and Thurmond, M. C. (2002). Predictive influence in the accelerated failure time model. *Biostatistics*, vol. 3 pp. 331-346.
- Burnham, K. P., and Anderson, D. R. (2002) *Model Selection and Inference (A Practical Information-Theoretic Approach)*. 2nd ed, New York: Springer.
- Collett, D. (2003). *Modeling Survival Data in Medical Research*, 2nd ed, New York: Chapman & Hall/CRC.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, vol. 34 pp. 187-202.
- Edmunson, J. H. Fleming, T. R. Decker, D. G. Malkasian, G. D. Jorgenson, E. O. Jeffries, J. A. Webb, M. J. and Kvols, L. K. (1979). Different Chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. *Cancer treatment reports*, vol. 63 pp. 241-247.
- Ghosal, S. and Ghosh, S. (2005). Semiparametric accelerated failure time models for censored data. *Bayesian Statistics and its Applications*, edited by S. K. Upadhyay, etc., pp. 15-39.

- Hanson T. and Johnson, W. O. (2004). A Bayesian semiparametric AFT model for interval-censored data. *Journal of Computational and Graphical Statistics*, vol. 13, pp. 341-361.
- Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, vol. 76 pp. 297-307.
- Hurvich, C. M. and Tsai, C.-L. (1990). Model selection for least absolute deviations regression in small samples. *Statist. Probab. Lett.*, vol. 9 pp. 259-265.
- Jin, Z. Lin, D. Y. Wei, L. J. and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, vol. 90, pp. 341-353.
- Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed, New York: Wiley.
- Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis (Techniques for Censored and Truncated Data)*, 2nd ed, New York: Springer.
- Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions (Theory and Applications)*, London: Imperial College.
- Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*, New York: Wiley.
- Lawless, J. F. (1980). Inference in the generalized gamma and log gamma distributions. *Technometrics*, vol. 22 pp. 409-419.
- Lindsey, J. K. (1996). *Parametric Statistical Inference*, New York: Oxford.
- Linhart, H. and Zucchini, W. (1986). *Model Selection*, New York: Wiley.
- McQuarrie, A. D. R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*, Singapore: World Scientific.
- Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, vol. 68 pp. 45-54.
- Robins, J. and Tsiatis, A. A. (1992). Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika*, vol. 79 pp.311-319.
- Robins, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, vol. 79 pp.321-334.
- Therneau, T. M. (1986). The COXREG Procedure. In *SAS SUGI Supplemental Library User's Guide*, fifth ed., North Carolina: SAS Institute Inc.
- Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for

- censored survival models. *Biometrics*, vol. 56 pp. 256-262.
- Walker, S. and Mallick, B. K. (1999). A Bayesian semiparametric accelerated failure time model. *Biometric*, vol. 55 pp. 477-483.
- Young, D. and Bakir, S. T. (1987). Bias correction for a generalized log-gamma regression model. *Technometrics*, vol. 29 pp. 183-191.