RIO DE JANEIRO
iSi 2015
60th World Statistics Congress

Model selection curves for survival analysis with accelerated failure time models

Karami, J. H.*, Luo, K., Fung, T.
Macquarie University, Sydney, Australia – md.karami@mq.edu.au,
kehui.luo@mq.edu.au, thomas.fung@mq.edu.au

Abstract

Many model selection processes involve minimizing a loss function of the data over a set of models. A recent introduced approach is model selection curves, in which the selection criterion is expressed as a function of penalty multiplier in a linear (mixed) model or generalized linear model. In this article, we have adopted *the model selection curves* for accelerated failure time (AFT) models of survival data. In our simulation study, it was found that for data with small sample size and high proportion of censoring, the model selection curves approach outperformed the traditional model selection criteria, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). In other situations with respect to sample size and proportion of censoring, the model selection curves correctly identify the true model whether it is a full or reduced model. Moreover, through bootstrapping, it was shown that the model selection curves can be used to enhance the model selection process in AFT models.

Keywords: log-likelihood function; penalty function; penalty multiplier; survival data.

1. Introduction

Many model selection strategies are based on minimizing $loss + penalty$ over a set of models. The penalty term usually consists of a penalty multiplier λ and a penalty function $f_n(p_\alpha)$, where p_α is the number of parameters in model α and n is the sample size. For a fixed penalty function, say $f_n(p_\alpha) = p_\alpha$, if we restrict ourselves to a single value of λ, we get a specififc model selection criteria. For example, $\lambda = 2$ will give AIC (Akaike, 1974) and $\lambda = \log(n)$ will give BIC (Schwarz, 1978). Other model selection criteria can be obtained using other values of λ, or other penalty functions. Recently, Müller and Welsh (2010) studied model selection criteria as a function of penalty multiplier, which resulted in what is known as *model selection curves*. In their study the *model selection curves* approach for linear regression models was illustrated, where they used residual sum of squares as the loss function. This approach allows researchers to study and rank candidate models before selecting a model. The stability of the selected model can be assessed in a range of λ values using the *model selection curves*. This approach has the potential to outperform other model selection criteria that are based on single value for penalty multiplier.

For many models, including accelerated failure time (AFT) model in survival analysis, it is common to use the log-likelihood to compare models. See for example Liang and Zou (2008), and Murray et al (2012). Our article concentrates on using the log-likelihood (l) for measuring the descriptive ability of an AFT model, and thus log-likelihood is a natural choice of loss function. There are also other loss functions, such as mean squared error, that can be used.

Suppose, we wish to choose one model α from all candidate models in $\mathcal{A}$ using *model selection curves*. The function of $loss + penalty$ can be expressed as:

$$\mathcal{M}(\lambda; \alpha) = -2l + \lambda p_\alpha, \qquad \lambda \geq 0, \alpha \in \mathcal{A} \tag{1}$$

For each specified λ, a model is chosen, which minimizes $\mathcal{M}(\lambda; \alpha)$ over $\alpha \in \mathcal{A}$ in (1). Now we can define a rank function:

$$r(\lambda; \alpha) = rank\left(\mathcal{M}(\lambda; \alpha)\right) \tag{2}$$

The $r(\lambda; \alpha)$ is computed at each λ. Note that rank functions are step functions, pairs of which have jumps at the values of λ where the ranks of models change. Now, assuming that no ties of $\mathcal{M}(\lambda; \alpha)$ occur for $\alpha \in \mathcal{A}$, the k $(1 \leq k \leq m$, and m is the number of models in $\mathcal{A})$ rank model selection curves can be defined by

$$C_{(k)}(\lambda; \mathcal{A}) = max\{\mathcal{M}(\lambda; \alpha);\ \ \alpha \in \mathcal{A}\ \wedge\ r(\lambda, \alpha) \leq k\} \tag{3}$$

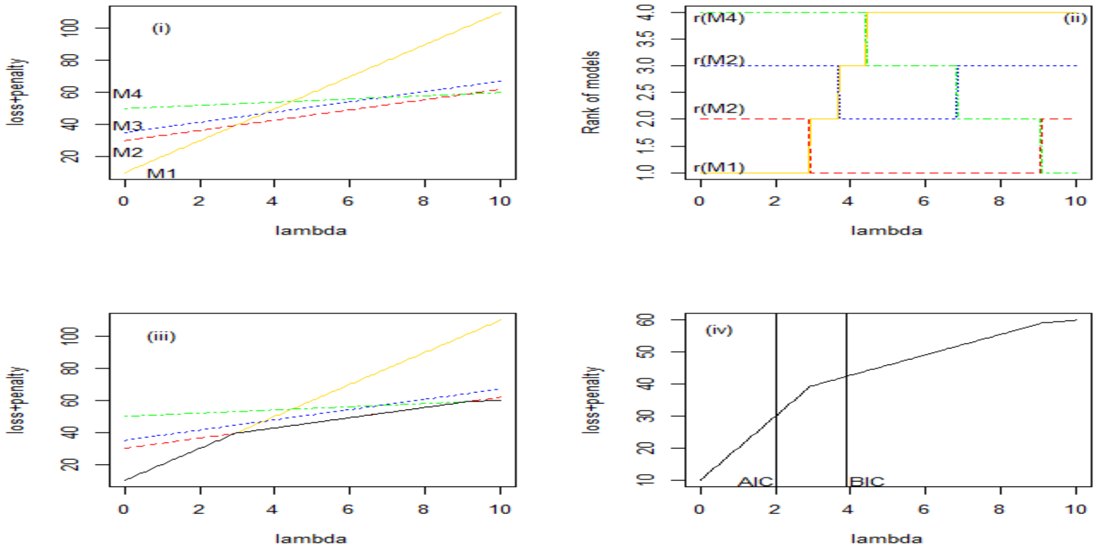
This definition can be extended to tied $\mathcal{M}(\lambda; \alpha)$'s for $\alpha \in \mathcal{A}$ by considering continuous locus of $C_{(k)}(\lambda; \mathcal{A})$ over $\lambda > 0$. The 1 rank model selection curve is the lower enveloping curve, which is defined as

$$C_{(1)}(\lambda; \mathcal{A}) = min\{\mathcal{M}(\lambda; \alpha);\ \ \alpha \in \mathcal{A}\ \} \tag{4}$$

This will be referred to as *model selection curve*. If perpendiculars are drawn from the vertices of the lower enveloping curve to the horizontal axis, the line segments on the horizontal axis give the respective length of cathetus corresponding to models that appear on the *model selection curves*. Thus model α should be chosen or selected if it has the longest cathetus in $C_{(1)}(\lambda; \mathcal{A})$, ie, the 1 rank *model selection curve*.

Let's now illustrate how to construct and use model selection curves using a simple example considering only four models, ie, $\alpha\ = 1, 2, 3, 4$. The plot of $\mathcal{M}(\lambda; \alpha)$ in $equation$ (1) against λ are shown on the top left of Figure 1 where $M1$ to $M4$ corresponds to model $\alpha = 1$ to 4, respectively. Plot of ranks, $r(\lambda; \alpha)$ in (2), against λ is on top right, ie, Figure 1(ii). Figure 1(iii) is the four rank selection curves, based on (3), and the model selection curve, obtained via (4), is presented on the bottom right of Figure 1. According to the model selection curve, Figure 1(iv), and also with reference to Figure 1(ii), model 2 (ie, $M2$) has the longest cathetus and is chosen from the four models considered.

Figure 1: Illustration of the construction of model selection curves



In this paper, it is intended to apply *model selection curves* for accelerated failure time models of right censored survival data using a simulation study. The rest of the paper is organized as follows: In Section 2, we briefly discuss the model selection curves in the case of Weibull AFT models. In Section 3, a simulation study with different combination of coefficients, censoring proportions and sample sizes are considered and discussed. In Section 4, we conclude our findings.

2. Model selection curves in Weibull AFT model

A general form of an AFT model is given by:

$$logT = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p + \tau W \tag{5}$$

where T is survival time with a specific distribution, W is random error term with a known form of density (eg, extreme value distribution), x_j's are predictors and α_j's are coefficients of x_j $(j = 1,\ 2, \cdots, p)$, α_0 is the model intercept and $\tau(\tau > 0)$ is a scale parameter.

Let $\{(y_i,\ \delta_i),\ i = 1,\ 2,\ \cdots, n\}$ denote the survival data with right censoring where $y_i = min(T_i,\ C_i)$, C_i is the censoring time for ith individual and $\delta_i = I(T_i \leq C_i)$ is the censoring indicator. Assuming the pairs $(y_i,\ \delta_i)$, $i = 1,\ 2,\ \cdots, n$ are independent, a general form of likelihood function for the survival data is $L = \prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}$, and thus the log-likelihood function is

$$l(\boldsymbol{\theta}) = \sum\nolimits_{i=1}^{n} \{\delta_i logf(y_i; x_i) + (1-\delta_i) logS(y_i)\}$$

where, $\boldsymbol{\theta} = (\alpha_0,\ \alpha^T,\ \tau)^T$, $\alpha^T = (\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_p)$, $f(\cdot)$ is the density function and $S(\cdot)$ is the survival function. The $l(\boldsymbol{\theta})$ can be expressed in terms of the density function ($g_0(w)$) and the survival function ($G_0(w)$) of W as below:

$$l(\boldsymbol{\theta}) = \sum\nolimits_{i=1}^{n} \left[\delta_i log\left\{\frac{1}{y_i\tau} g_0(w_i)\right\} + (1-\delta_i) logG_0(w_i)\right]$$

or,

$$l(\boldsymbol{\theta}) = -log\tau \sum\nolimits_{i=1}^{n} \delta_i + \sum\nolimits_{i=1}^{n} \{\delta_i logg_0(w_i) + (1-\delta_i) logG_0(w_i)\} - \sum\nolimits_{i=1}^{n} \delta_i logy_i \qquad (6)$$

where, $w_i = \frac{logy_i - \alpha_0 - \alpha^T x_i}{\tau}$.

A Weibull AFT model arises if the error term W in model (5) follows an extreme value distribution with density and survival functions respectively as $g_0(w) = exp[w - exp(w)]$ and $G_0(w) = exp[-exp(w)]$, $w \in \mathbb{R}$. Plugging these two functions into equation (6) gives the log-likelihood for the Weibull AFT model as:

$$l(\boldsymbol{\theta}) = \left(\frac{1-\tau}{\tau}\right) \sum\nolimits_{i=1}^{n} \delta_i logy_i - \sum\nolimits_{i=1}^{n} exp\left(\frac{logy_i - \alpha_0 - \alpha^T x_i}{\tau}\right) - \sum\nolimits_{i=1}^{n} \delta_i \left(log\tau + \frac{\alpha_0}{\tau} + \alpha^T x_i\right)$$

The maximum likelihood estimates (MLE) of the parameters can be obtained from the above equation using an iterative procedure (eg, Newton-Raphson), and the model selection curves for the Weibull AFT model can then be constructed from

$$\mathcal{M}(\lambda; \alpha) = -2l(\widehat{\boldsymbol{\theta}}) + \lambda p_\alpha,\ \ \lambda > 0 \qquad (7)$$

where, $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$.

As expected, both AIC and BIC would appear as single points on the *model selection curves*. To construct *model selection curves*, we choose $\lambda\ \in [0,\ 4log(n)]$ in (7), same as Müller and Welsh's study (2010). Such range of λ would cover most of existed model selection criteria that are based on single values of $\lambda, such\ as\ AIC\ and\ BIC.$

3. A simulation study

For our simulation study, we consider Weibull AFT model with four predictors, ie,

$$logT_i = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \tau W_i, \qquad i = 1,\ 2,\ \cdots, n.$$

Observations for the (weakly correlated) predictors $x_1,\ x_2,\ x_3,\ x_4$ are drawn from a multivariate normal distribution $MN(\mathbf{0}, \boldsymbol{\Sigma}_0)$ with zero mean vector and dispersion matrix, $\boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & 0.001 & 0.001 & 0.001 \\ 0.001 & 1 & 0.001 & 0.001 \\ 0.001 & 0.001 & 1 & 0.001 \\ 0.001 & 0.001 & 0.001 & 1 \end{bmatrix}$. We considered different sample sizes ranging from 30 to 300 at various censoring proportions such as 10% and 50% along with different sets of coefficients: coefs1 $= (0.1,\ 0,\ 0,\ 0.9,\ 0.8)$, coefs2 $= (0.1,\ 0,\ 0,\ 0.9,\ 0)$, and coefs3 $= (0.1,\ 1,\ 0.7,\ 0.9,\ 0.8)$. Both uncensored and censored observations (ie, T_i and C_i) are drawn separately from Weibull distribution with specified shape and scale. Then survival times are defined as Time $= min(T_i,\ C_i)$. Note that the survival times are related with predictors through the scale parameter of Weibull distribution.

Here are some of the results from our simulation study. Rank plots, ie, $r(\lambda, \alpha)$ against λ, for different combinations of coefficients, sample sizes and censoring proportions are presented in Figures 2 – 5. It is clear from Figure 2 that, when the full model (indicated by coefs3) is true, the model selection curves and other model selection criteria, such as AIC and BIC, can correctly identify it, irrespective of sample sizes and even at high censoring proportion. If a reduced model $\{1,\ 4,\ 5\}$ (ie, model with predictors x_3 and x_4) or model $\{1,\ 4\}$ (ie, model with predictor x_3) is true, the model selection curves and other model selection criteria also identify the true model correctly if sample size is large ($n = 150, 300$) and censoring proportion (50%) is high as shown in Figure 3.

Figure 2: Rank plots of full model
Censoring proportion=0.50, coefs3=(0.1, 1, 0.7, 0.9, 0.8)

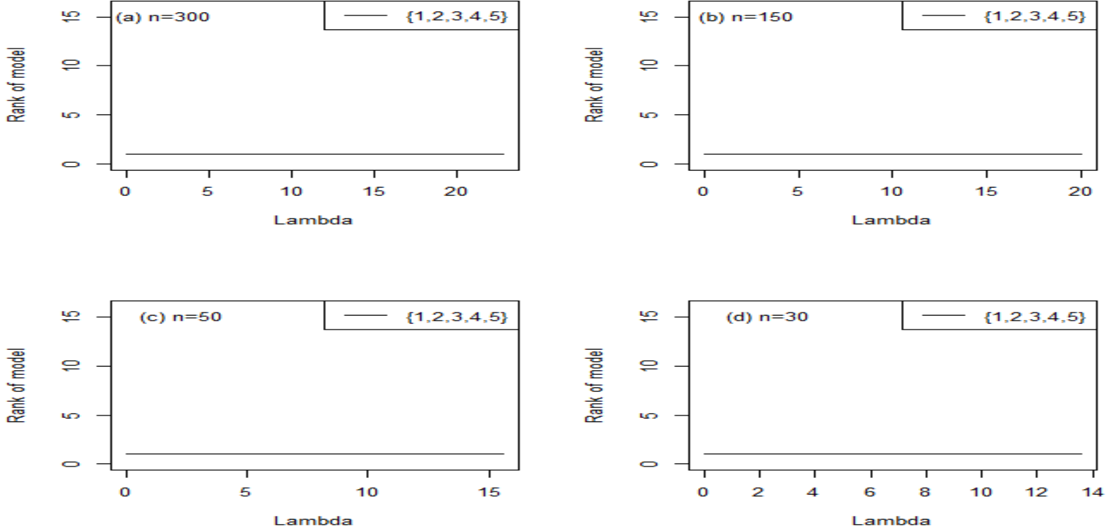


Figure 3: Rank plots of submodel with high censoring (50%)
coefs1=(0.1, 0, 0, 0.9, 0.8), coefs2=(0.1, 0, 0, 0.9, 0)

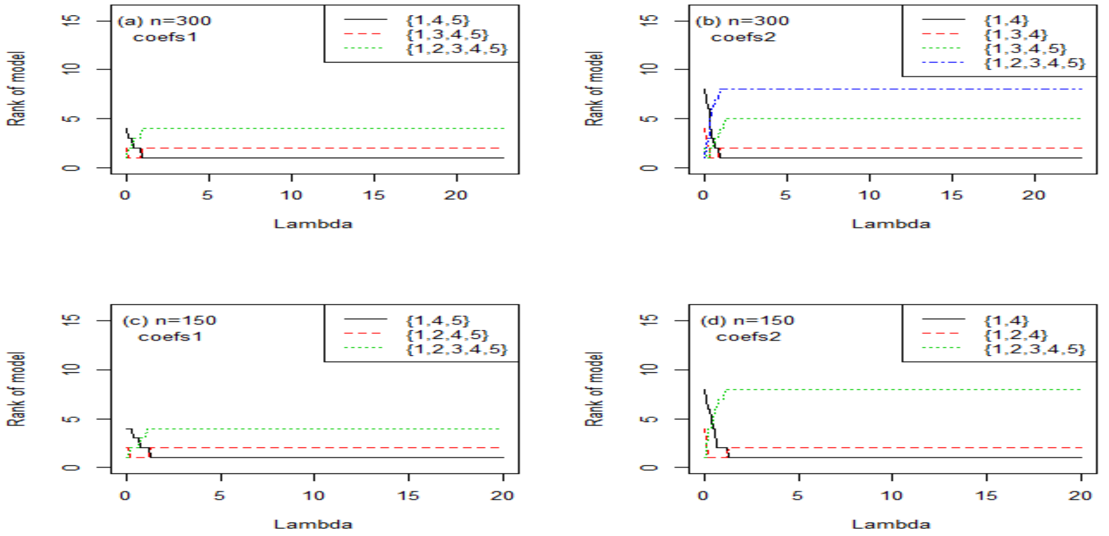


Figure 4 shows the rank plots along with the catheti for models α that appear on the model selection curves for data with sample sizes $n = 30, 50$ and a low censoring proportion of 10%. In all cases, the model selection curves can correctly identify the true model (coefs1 in (a) and (c) or coefs2 in (b) and (d)) while AIC and BIC cannot.

It is possible that two or more models may have similar length of cathetus on the model selection curves. In this case, we may use bootstrap method to provide additional information to model selection process. Bootstrapping can also be used to quantify the importance of each variable considered by computing the proportion of times a predictor is included in a model at each λ on the model selection curves. See for, instance, Figure 5(d). We may also estimate the chance of selecting a particular model from the model selection curves. The empirical bootstrap estimate $\pi^*(\lambda; \alpha)$ of the probability of selecting model α is obtained by computing proportion of models at each λ that appear on the model selection curves. The average of these proportions is a bootstrap estimate $\pi^*(\alpha)$ of the probability of selecting model α. In our simulation study, we have used 1,000 bootstrap replications.

Figure 4: Rank plots of submodel with low censoring (10%) and small sample; coefs1=(0.1, 0, 0, 0.9, 0.8), coefs2=(0.1, 0, 0, 0.9, 0)

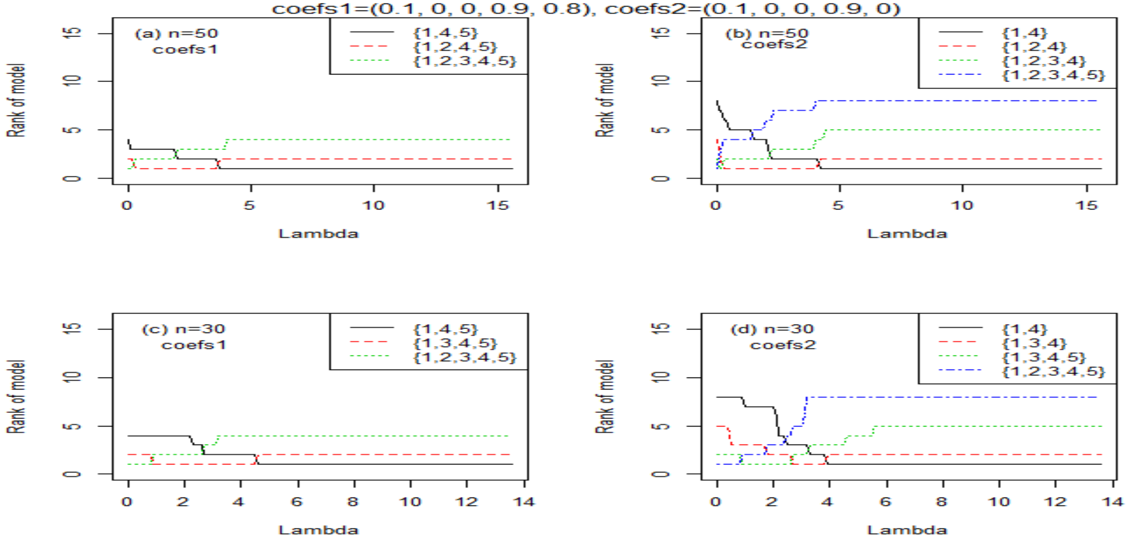
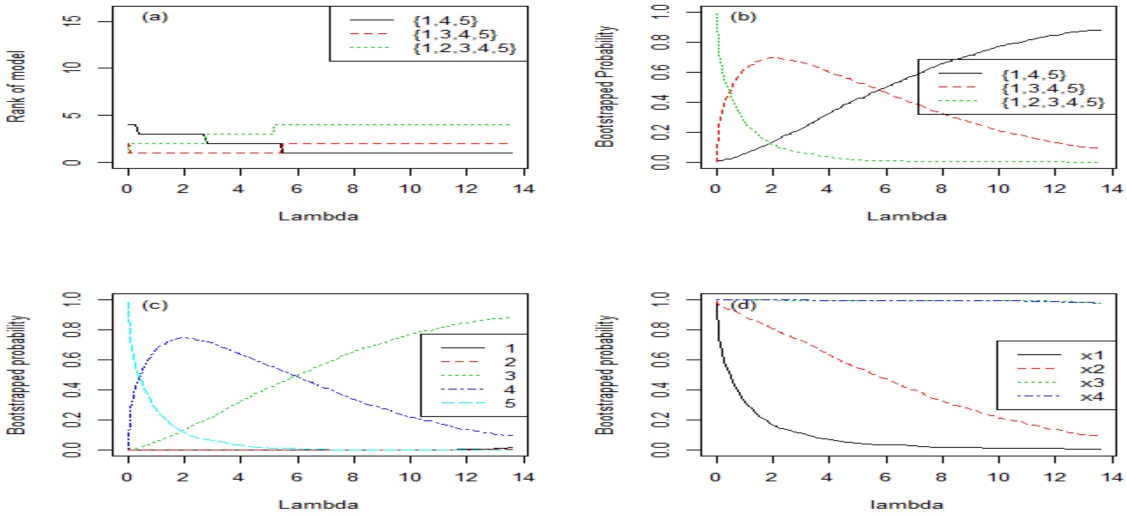


Figure 5: (a) Rank plot, (b) Single model plot, (c) Dimensionality plot, (d) Variable inclusion plot



In Figure 5, we have considered data from coefs1 with sample size of 30 and a high censoring proportion of 50%. In this case, both AIC and BIC select model $\{1,3,4,5\}$ when the true model is $\{1,4,5\}$ as indicated by coefs1. However, the model selection curves identify the true model $\{1,4,5\}$ as shown in Figure 5(a) where this model has the longest cathetus although it is not that much longer than that for model $\{1,3,4,5\}$. In such case, one may use bootstrapping for more information. Figure 5(b) shows the plot of bootstrap probabilities, $\pi^*(\alpha)$ against λ for models that appear on the *model selection curves* with selection probability larger than 4%. It is clear that $\pi^*(\alpha)$ for the true model steadily increases in the entire range of λ. From Figure 5(c), it is seen that a model with dimension 3 (ie, containing 3 parameters) may have a greater likelihood. Figure 5(d) shows that both predictors x_3, x_4 have excellent description and prediction qualities because their bootstrapped probabilities are almost always 1 over the entire range of λ values. All these support the model identified by the model selection curve, ie, model $\{1,4,5\}$. If we wish to consider one more predictors, x_2 is certainly more useful than x_1, as shown in Figure 5(d). We have only reported some of the results from our simulation study here, but more detailed results are available from the authors.

4. Conclusions

We have seen that if the full model is true, model selection curves and other selection criteria usually identify it correctly irrespective of sample size and censoring proportion (up to 50%). In the case of a sub-model (ie, model not including all predictors considered) is true, similar results are also observed for large sample sizes (say, 300) at high censoring proportion such as 50%. If a sub-model is true but the sample size is small with a low censoring proportion (eg, 10%), the model selection curves still select the correct model while AIC and BIC may not. Even in the case that the sub-model is true, for data with small size (say, 30) and high proportion of censoring (say, 50%), bootstrapped results show that the model selection curves are more likely to identify the correct model than other criteria such as, AIC and BIC. Thus we conclude that the model selection curves approach may outperform other model selection criteria in some instances when selecting Weibull AFT models.

References

Akaike, H. (1973). Information theory and an extension of the maximum liklihood principle. In Proceedings of the Second International Symposium of Information Theory, Eds. B.N. Petrov, F. Csáki, pp. 267 – 281. Akadémiai Kiadó: Budapest.

Liang, H. and Zou, G. (2002). Improved AIC selection strategy for survival analysis. CS&DA, 52(5): 2538–2548.

Müller, S. and Welsh, A.H. (2010). On Model Selection Curves. International Statistical Review, 78(2): 240–256.

Murray, K., Heritier, S. and Müller, S. (2013). Graphical tools for model selection in generalized linear models. Statistics in Medicine, 32, 4438 – 4451.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of statistics, 6(2): 461–464.