Assessing the adequacy of Weibull survival models: a simulated envelope approach

Yun Zhao a, Andy H. Lee a, Kelvin K.W. Yau b & Geoffrey J. McLachlan c

a School of Public Health, Curtin Health Innovation Research Institute, Curtin University of Technology, Perth, WA, Australia
b Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China
c Department of Mathematics, University of Queensland, Brisbane, Queensland, Australia

Available online: 15 Jan 2011

To cite this article: Yun Zhao, Andy H. Lee, Kelvin K.W. Yau & Geoffrey J. McLachlan (2011): Assessing the adequacy of Weibull survival models: a simulated envelope approach, Journal of Applied Statistics, 38:10, 2089-2097

To link to this article: http://dx.doi.org/10.1080/02664763.2010.545115

Please scroll down for article

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan, sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Assessing the adequacy of Weibull survival models: a simulated envelope approach

Yun Zhao\textsuperscript{a}, Andy H. Lee\textsuperscript{a,}\textsuperscript{*}, Kelvin K.W. Yau\textsuperscript{b} and Geoffrey J. McLachlan\textsuperscript{c}

\textsuperscript{a}School of Public Health, Curtin Health Innovation Research Institute, Curtin University of Technology, Perth, WA, Australia; \textsuperscript{b}Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China; \textsuperscript{c}Department of Mathematics, University of Queensland, Brisbane, Queensland, Australia

(Received 10 March 2010; final version received 9 November 2010)

The Weibull proportional hazards model is commonly used for analysing survival data. However, formal tests of model adequacy are still lacking. It is well known that residual-based goodness-of-fit measures are inappropriate for censored data. In this paper, a graphical diagnostic plot of Cox–Snell residuals with a simulated envelope added is proposed to assess the adequacy of Weibull survival models. Both single component and two-component mixture models with random effects are considered for recurrent failure time data. The effectiveness of the diagnostic method is illustrated using simulated data sets and data on recurrent urinary tract infections of elderly women.

Keywords: goodness-of-fit; mixture models; model adequacy; simulated envelope; survival analysis

1. Introduction

The Weibull proportional hazards model is one of the most commonly used parametric models in survival analysis. An important issue is to determine the adequacy of the fitted survival model. In the literature, attention has been focused on both graphical and analytical goodness-of-fit techniques for the Cox’s proportional hazards model \cite{1,6,11,12,14,17}. These diagnostic methods are mostly derived from the martingale residual or its variants, and mainly developed for assessing the proportional hazards assumption, whereas measures analogous to $R^2$ in multiple regression are not readily available. Indeed, a single, simple and easy to interpret measure for the Cox regression model does not exist \cite[p. 229]{7}. The main reason is due to the presence of censored observations in the survival data. A low $R^2$ type value is likely even for an adequate model when the proportion of censoring is high. Likewise, assessing model adequacy based on the standard Cox–Snell residual plot is not straightforward, because the specific departure(s) may not be apparent when the survival model is incorrectly specified \cite[p. 128]{8}. According to Therneau and Grambsch \cite[pp. 81–83]{16}, the overall distribution of the martingale residuals does not aid

\textsuperscript{*}Corresponding author. Email: andy.lee@curtin.edu.au

ISSN 0266-4763 print/ISSN 1360-0532 online
© 2011 Taylor & Francis
DOI: 10.1080/02664763.2010.545115
http://www.informaworld.com
in the global assessment of fit, while the best model need not have the smallest sum of squared martingale residuals. The sum of squared deviance residuals also does not necessarily decrease with improved model fit. Consequently, the sum of squared residuals generally cannot be used as a summary measure to determine the overall goodness-of-fit of survival models [16, p. 81]. In the related context of sensitivity analysis, however, influence diagnostics have been developed to identify anomalous observations for survival models [19].

The aim of this paper is to develop a suitable diagnostic method for assessing the adequacy of Weibull proportional hazards models. In Section 2, the mathematical framework underlying Weibull survival mixture models is briefly reviewed. For recurrent failure time data, random effects (frailties) are incorporated within the model to account for the inherent dependency of the observations. In Section 3, an assessment procedure using Cox–Snell residuals is constructed based on a simulated envelope approach. A simulation study is conducted in Section 4 to investigate the validity of the diagnostic procedure. The technique is further illustrated in Section 5 with an example on recurrent urinary tract infections (UTI) of elderly women. Finally, some concluding remarks are provided in Section 6.

2. Weibull proportional hazards models

Consider recurrent failure time data from $M$ patients with a particular disease. Let $T_{ij}$ be the observable failure/censoring time for the $j$th failure observation ($j = 1, 2, \ldots, n_i$) for patient $i (i = 1, 2, \ldots, M)$, and $N = \sum_{i=1}^{M} n_i$ being the total number of observations. The observed censoring indicator is denoted by the binary variable $\delta_{ij}$ as usual. Let $x_{ij}$ represent a vector of covariates associated with $T_{ij}$. Under the proportional hazards assumption, the hazard function of the $j$th recurrent failure event for patient $i$ at time $t$ is given by

$$h(t_{ij}; x_{ij}) = h_0(t_{ij}) \exp(\eta(x_{ij})), \quad (1)$$

where $h_0(t_{ij})$ is the baseline hazard function and $\eta(x_{ij}) = x_{ij}^T \beta + U_i$, where $\beta$ is the vector of regression coefficients associated with covariate $x_{ij}$, and where $U_i \sim N(0, \theta)$ denotes the unobserved random effect of patient $i$. The normal distribution is assumed for the random effects in view of its established inferences in linear mixed models and readiness for assigning correlation in the covariance matrix.

Assuming a Weibull distribution for the baseline hazard function $h_0(t_i)$,

$$h_0(t_{ij}) = \lambda \gamma t_{ij}^{\gamma-1}, \quad (2)$$

where $\lambda, \gamma > 0$ are unknown scale and shape parameters, respectively. The survival function corresponding to Equation (2) is then $S(t_{ij}; x_{ij}) = \exp(-\lambda \exp(\eta(x_{ij}))t_{ij}^\gamma)$. The fitting of the Weibull proportional hazards model can be implemented by consideration of the joint log-likelihood function $L$ based on the observed data and the random effects terms in the model [19]:

$$L = \sum_{i=1}^{M} \sum_{j=1}^{n_i} [\delta_{ij} \log f(t_{ij}; x_{ij}) + (1 - \delta_{ij}) \log S(t_{ij}; x_{ij})] - \frac{1}{2} M \log(2\pi \theta) + \frac{u^T u}{\theta}, \quad (3)$$

where $f(t_{ij}; x_{ij})$ is the Weibull probability density function and $u^T = [U_1, U_2, \ldots, U_M]$. A two-stage estimation procedure can be adopted, whereby for a specified value of $\theta$ in the distribution of the random effect terms, this log-likelihood function is maximized over the other parameters and the values of the random effect terms. Then with the latter held fixed at their currently maximized values, the vector of unknown parameters is estimated by adopting a residual maximum likelihood (REML) approach [19]. A numerical procedure, involving Newton–Raphson coupled with the EM algorithm, has been suggested to implement this two-stage estimation process [13].
When the recurrent events are characterized by an acute phase followed by a stable phase after the index episode, a two-component Weibull survival mixture model is deemed appropriate for such heterogeneous failure time data with two overlapping phases [10,13]. Under this setting, the survival function of $T$ is modelled as

$$S(t_{ij}; x_{ij}) = pS_1(t_{ij}; x_{ij}) + (1 - p)S_2(t_{ij}; x_{ij}),$$

(4)

with $h_g(t_{ij}; x_{ij}) = h_{g0}(t_{ij}) \exp(\eta_g(x_{ij}))$, where $p$ denotes the proportion of observations in the first component, $h_{g0}(t_{ij})$ is the baseline hazard function commonly taken to be the Weibull distribution:

$$h_{g0}(t_{ij}) = \lambda g \gamma g (t_{ij})^{\gamma g - 1};$$

$\eta_g(x_{ij})$ is the linear predictor relating to the covariates $x_{ij}$; $S_g(t_{ij}; x_{ij})$ and $h_g(x_{ij})$ are the conditional survival function and hazard function of the $g$th component ($g = 1, 2$), respectively. The linear predictor, with unobserved random effects $U_{gi}$ ($g = 1, 2$), is specified as

$$\eta_g(x_{ij}) = x_{ij}^T \beta_g + U_{gi},$$

(5)

where $\beta_g$ is the vector of regression coefficients and the random effects $U_{gi}$ are taken to be i.i.d. $N(0, \theta_g)$ [10]. For a given initial value of $\theta_g$, the joint log-likelihood function to be maximized is $L = L_1 + L_2$, where

$$L_1 = \sum_{i=1}^{M} \sum_{j=1}^{n_i} [\delta_{ij} \log f(t_{ij}; x_{ij}) + (1 - \delta_{ij}) \log S(t_{ij}; x_{ij})],$$

$$L_2 = -\frac{1}{2} [M \log 2\pi \theta_1 + (1/\theta_1)u_1^T u_1] - \frac{1}{2} [M \log 2\pi \theta_2 + (1/\theta_2)u_2^T u_2].$$

(6)

Here, $L_1$ represents the log-likelihood of recurrent times conditional on random effects vectors $u_1$ and $u_2$, whereas $L_2$ is the logarithm of the joint probability density function of $u_1$ and $u_2$, with $u_1^T = [U_{11}, U_{12}, \ldots, U_{1M}]$ and $u_2^T = [U_{21}, U_{22}, \ldots, U_{2M}]$ being independent. An approximate REML estimator for the variance component parameters can be obtained; see Refs [10,13] for more details.

3. Assessment of model adequacy

For linear regression models, Atkinson [2,3] proposed a half-normal diagnostic plot to detect potential outliers and influential observations, in which the ordered statistics of the absolute deletion residuals are plotted against the ordered quantiles $z(k + N - 1/8) / (2N + 1/2)$, where $N$ is the number of observations and $z(\alpha)$ denotes the $\alpha$-percentile of the standard normal distribution. A simulated envelope, generated from the fitted regression model, is added to the plot to aid overall assessment, whereby the observed residuals are expected to lie within the boundary of the envelope if the presumed model has been correctly specified. The main advantage of this simulation technique is its ease of interpretation without imposing any assumption on the residual distribution. The method has been successfully applied to confirm multiple outliers in generalized linear and nonlinear regressions [9]. Recently, simulated envelope plots of Pearson residuals are also used to diagnose overdispersion in count data models [18].

The principle underlying the simulated envelope technique is to compare the observed statistics with those of the surrogated data generated from the hypothesized model. The method is particularly useful for failure time data in view of the difficulties associated with survival model assessment, especially since the distribution of the residuals is rarely known. Any systematic departure of the observed residuals from the simulated quantities, with reference to the boundary of the envelope, may be taken as evidence against the adequacy of the assumed model.
The diagnostic procedure consists of the following steps:

1. Fit the Weibull proportional hazards model to the observed failure time data.
2. Generate a sample of $N$ observations based on the parameter estimates obtained from the fitted model. Censoring or event is determined by whether the generated failure time exceeds a specified constant $C$ so that the resulting censorship rate is similar to that of the observed data set.
3. Fit the specified Weibull model to the replicated sample above, and compute the ordered values of the Cox–Snell residuals, which are defined as
   \[ e_{ij} = -\log \hat{S}(t_{ij}; x_{ij})|_{\hat{\psi}}, \quad i = 1, 2, \ldots, M, \quad j = 1, 2, \ldots, n_i, \]
   where $\hat{S}(t_{ij}; x_{ij})|_{\hat{\psi}}$ is the estimated value of the cumulative survival function evaluated at the estimate $\hat{\psi}$ of the unknown parameter vector $\psi$.
4. Repeat steps 2 and 3 $H$ number of times.
5. Consider the $N$ sets of the $H$ ordered statistics of the Cox–Snell residuals; calculate the average, minimum and maximum values across each set.
6. Plot these values together with the ordered residuals from the original data in an index plot. The minimum and maximum values of the $H$ ordered statistics constitute a simulated envelope to guide assessment of the model adequacy.

Atkinson [2, p. 36] originally suggested using $H = 19$ for a 5% chance to detect the largest residual being outside the boundary of the simulated envelope if the fitted regression model was correctly specified. The capability of modern computers enables the simulation of a large number of replicated data sets, and $H = 200$ is recommended for the current survival data setting. Moreover, other types of residuals such as martingale, deviance or score residual may be used in the above diagnostic procedure instead of the Cox–Snell residual.

4. Simulation study

A simulation study is conducted to investigate the performance of the proposed diagnostic procedure, by comparing the (true) two-component Weibull mixture model versus the (misspecified) single-component Weibull model. Let $M = 20$ and $n_i = 25$. Two covariates are considered, with $x_1$ being generated as independent $N(0, 1)$ random variables and $x_2$ is generated from a Bernoulli distribution with probability 0.5. Under the mixture model setting, $N = 500$ realizations of an unobservable variable $Y$ are generated such that an individual has probability $p$ and $(1 - p)$ being belonging to the first and second component, respectively.

For the first component, failure times are generated based on the Weibull proportional hazards model defined in Section 2, with $\lambda = 0.05$, $\gamma = 1.5$ and $\eta_{1i} = 0.5x_{1i} - 0.5x_{2i} + U_{1i}$, where $U_{1i}$ is generated from $N(0, \theta_1)$. Similarly, failure times for the second component are also generated based on the Weibull proportional hazards model, but with $\lambda = 0.01$, $\gamma = 0.5$ and $\eta_{2i} = x_{1i} - x_{2i} + U_{2i}$, where $U_{2i}$ is generated from $N(0, \theta_2)$. Censoring or event is determined by whether the generated failure time exceeds a constant $C$, which is set to be 1000 in this simulation study. By varying $p$ (= 0.5, 0.75, 0.9) and $\theta_1 = \theta_2(=0.5, 1)$, a total of six surrogated data settings are considered.

Figure 1 presents the simulated envelope plots of Cox–Snell residuals from fitting a Weibull proportional hazards model with random effects to the six data sets. Most of the observed Cox–Snell residuals are outside the boundary of the envelope, indicating inadequacy of the misspecified single-component model. On the other hand, the observed sets of Cox–Snell residuals corresponding to the two-component Weibull model are well within the respective simulated envelope.
Figure 1. Simulated envelope plot of Cox–Snell residuals from fitting a Weibull proportional hazards model with random effects to the six surrogated data sets: (a) \( p = 0.5, \theta_1 = 0.5, \theta_2 = 0.5 \); (b) \( p = 0.5, \theta_1 = 1, \theta_2 = 1 \); (c) \( p = 0.75, \theta_1 = 0.5, \theta_2 = 0.5 \); (d) \( p = 0.75, \theta_1 = 1, \theta_2 = 1 \); (e) \( p = 0.9, \theta_1 = 0.5, \theta_2 = 0.5 \); (f) \( p = 0.9, \theta_1 = 1, \theta_2 = 1 \); where “plus symbols” denote the observed ordered residuals, the solid line represents the average residual values and the dashed lines show the minimum and maximum ordered residuals across 200 replications.

in Figure 2, confirming no evidence of lack of fit of the hypothesized survival mixture model. Therefore, the simulation results show that the proposed diagnostic procedure is valid for assessing the adequacy of Weibull survival models.

5. Application to recurrent urinary tract infections data

A retrospective cohort study was conducted to determine the risk factors associated with recurrent UTIs among elderly women in residential aged-care facilities [10]. Eligibility criteria for the subjects were defined to be female residents aged 60 years or above with an institutionalization period of at least six months. A total of 201 women were recruited from six aged-care institutions in Perth, Western Australia. The outcome variable was the recurrent time between successive UTI episodes. During the two-year follow-up period, 285 observations were recorded from \( M = 93 \) women, with the maximum number of recurrent UTI being 17. The average age of the cohort was 85.8 (S.D. 8.4) years and 32 (34%) of them had a history of prior UTI. The mean recurrence time was 241 (S.E. 19.6) days.
Patients with recurrent UTI can be categorized as being in acute or stable condition. These two phases typically overlap in time so that the risk of recurrence cannot be described satisfactorily by fitting separate parametric survival models to each phase alone. A two-component survival mixture model is deemed appropriate to model the recurrent times [10]. For comparison purpose, both single-component and two-component Weibull proportional hazards mixture models described in Section 2 are fitted to the recurrent UTI data, with covariates age and history of prior UTI taken at baseline. The model estimation results are summarized in Table 1. The results suggest that about 18% of the recurrent UTI among elderly women belong to the stable condition state. For the regression coefficients $\beta_1$ and $\beta_2$, the standard Wald statistic is used for assessing significance relative to the $\chi^2(1)$ reference distribution. For the mixture proportion and variance components, testing the null hypothesis $p = 0$ or $\theta = 0$ involves the boundary of the parameter space, so that the test statistic has a 50:50 mixture distribution of $\chi^2(0)$ and $\chi^2(1)$ [4,5,15].

According to the two-component survival mixture model, the hazard rate of recurrent UTI is significantly associated with the patient’s history of prior UTI during the acute phase (the proportion of which is estimated to be 82%). The random patient effects are significant in both acute...
Table 1. Parameter estimates (standard error) from fitting Weibull proportional hazards model and two-component Weibull proportional hazards mixture model to the recurrent UTI data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Weibull proportional hazards model</th>
<th>First component</th>
<th>Second component</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>0.818 (0.036)*</td>
<td>0.525 (0.215)*</td>
<td>0.521 (0.141)*</td>
</tr>
<tr>
<td>$\theta$</td>
<td>0.200 (0.095)*</td>
<td>1.146</td>
<td>2.401</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.939</td>
<td>-6.159</td>
<td>-11.610</td>
</tr>
<tr>
<td>$\log \lambda$</td>
<td>-4.650</td>
<td>-0.010 (0.016)</td>
<td>0.044 (0.036)</td>
</tr>
<tr>
<td>Age ($\beta_1$)</td>
<td>-0.009 (0.001)*</td>
<td>0.973 (0.276)*</td>
<td>-0.030 (0.356)</td>
</tr>
<tr>
<td>Prior UTI ($\beta_2$)</td>
<td>0.765 (0.169)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>2336.685</td>
<td></td>
<td>1004.12</td>
</tr>
<tr>
<td>BIC</td>
<td>2351.295</td>
<td></td>
<td>1036.99</td>
</tr>
</tbody>
</table>

Note: *p-value < 0.05.

and stable phases, implying that heterogeneity in UTI recurrence can be attributed to individual patient variations. Meanwhile, the effects of both age and prior UTI experience on the recurrent time are evident based on the single-component Weibull proportional hazards model. Akaike’s information criterion (AIC) and Bayesian information criterion (BIC) statistics are computed as indicative measures of model goodness-of-fit, both of which favour the mixture model over its single component counterpart for the UTI data.

The diagnostic procedure outlined in Section 3 is next implemented for both models. The simulated envelope plot corresponding to the Weibull proportional hazards model is shown in Figure 3. Most of the observed Cox–Snell residuals are either near or outside the boundary of the envelope, indicating inadequacy of the fitted single-component model. On the other hand, the

Figure 3. Simulated envelope plot of Cox–Snell residuals from fitting a Weibull proportional hazards model with random effects to the recurrent UTI data, where “plus symbols” denote the observed ordered residuals, the solid line represents the average residual values and the dashed lines show the minimum and maximum ordered residuals across 200 replications.
observed set of Cox–Snell residuals corresponding to the two-component Weibull proportional hazards model in Figure 4 are well within the simulated envelope, confirming no evidence of lack of fit of the hypothesized survival mixture model. Similar results are also obtained using the martingale residuals and thus not presented for brevity.

6. Concluding remarks

This paper presents a simulated envelope approach to assess the adequacy of Weibull survival models. A simulation study and application to a set of recurrent UTI data illustrate the usefulness of the graphical diagnostics, which suggest that the two-component Weibull proportional hazards mixture model can provide a reasonable fit to the data. However, the method cannot be treated as a formal testing procedure, since the exact probability of the envelope containing the sample is not known [2, p. 36].

Computer programs for fitting the Weibull proportional hazards models and constructing the simulated envelope have been implemented in the SPlus platform. Although the codes are specifically written for one- and two-component Weibull proportional hazards models in the repeated measurements setting, they can be modified to the standard situation without random effects. They can also be further developed into a stand alone package to assess different specifications of the parametric distribution and cater for other survival mixture models, such as the Weibull accelerated failure time model, the log-logistic model, and the Gamma model, among others [10].

The simulated envelope provides a simple and effective diagnostic method to assess the adequacy of survival models. However, there are two limitations of this graphical approach. Firstly, the procedure relies on some specification of the parametric distribution for the survival time. It is thus not applicable when no information is available concerning the underlying survival distribution. Secondly, the proportion of censoring may vary among the replicated samples. Consequently, the ratios of censored observations to realized failure time events are likely to be different between
the simulated data sets, which will impact on the accuracy of the resulting simulated envelope. In practice, one may conduct preliminary simulations to estimate the underlying censored time $C$ before running step 2 of the diagnostic procedure, so as to avoid substantial departures from the original censoring proportion.

Acknowledgements

This research was supported by grants from the National Health and Medical Research Council of Australia (ID 425510) and the Research Grants Council of Hong Kong.

References