STATISTICS IN MEDICINE
*Statist. Med.* 2002; **21**:3493–3510 (DOI: 10.1002/sim.1251)

# Comparing proportional hazards and accelerated failure time models for survival analysis

Jesus Orbe, Eva Ferreira and Vicente Núñez-Antón*,†

*Departamento de Econometría y Estadística Facultad de Ciencias Económicas y Empresariales, Universidad del País Vasco Euskal/Herriko Unibertsitatea, Bilbao, Spain*

SUMMARY

This paper describes a method proposed for a censored linear regression model that can be used in the context of survival analysis. The method has the important characteristic of allowing estimation and inference without knowing the distribution of the duration variable. Moreover, it does not need the assumption of proportional hazards. Therefore, it can be an interesting alternative to the Cox proportional hazards models when this assumption does not hold. In addition, implementation and interpretation of the results is simple. In order to analyse the performance of this methodology, we apply it to two real examples and we carry out a simulation study. We present its results together with those obtained with the traditional Cox model and AFT parametric models. The new proposal seems to lead to more precise results. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: lifetime; Kaplan–Meier weights; jack-knife; censored data; non-proportional hazards

## 1. INTRODUCTION

In survival analysis, in general we have censored observations and, because of this and other characteristics, such as, for example, its asymmetric distribution, the usual statistical methods or techniques cannot be applied to this type of data. As a consequence, we find in the statistical literature specific models for survival data or lifetime analysis. If we consider regression models, the ones most used are the Cox proportional hazards model [1] (PH) and the accelerated failure time models [2] (AFT).

* Correspondence to: Vicente Núñez-Antón, Departamento de Econometría y Estadística, Facultad de Ciencias Económicas y Empresariales, Avda. Lehendakari Agirre, 83 E-48015, Bilbao, Spain.
† E-mail: vn@alcib.bs.ehu.es

Contract/grant sponsor: Universidad del País Vasco/Euskal Herriko Unibertsitatea; contract/grant number: UPV 038.321-HA129/99, UPV 038.321-13631/2001
Contract/grant sponsor: Dirección General de Enseñanza Superior e Investigación Científica del Ministerio Español de Educación y Cultura; contract/grant number: PB98-0149
Contract/grant sponsor: Gobierno Vasco; contract/grant number: PI-1999-46, PI-1999-70

*Received January 2001*
*Accepted December 2001*

The first one and its various generalizations are mainly used in medical and biostatistical fields, while the alternative class of regression models, the AFT model, is mainly used in reliability theory and industrial experiments.

The most commonly used model is the PH model. Its use is preferred because estimation and inference about the parameters of interest are possible without assuming any form for the baseline hazard function, that is, it is not necessary to specify a survival distribution to model the effect of the explanatory variables on the duration variable. However, this model is based on the PH assumption and this may not hold in some survival studies. In a recent review of survival analyses in cancer journals [3], it was found that only 5 per cent of all studies using the Cox PH model attempted to verify the underlying assumption. If this assumption does not hold, the standard Cox model should not be used and may entail serious bias and loss of power when estimating or making inference about the effect of a given prognostic factor on mortality [4, 5]. Many techniques for assessing the goodness-of-fit of PH regression models and methods for detecting violations of this assumption can be found in the statistical literature [6–11]. In the 1990s several flexible methods were proposed to take into account the non-proportionality of hazards [4, 5, 11, 12].

On the other hand, if we consider the AFT models, these models could be of interest because they can be rewritten specifying a direct relation between the logarithm of the survival time and the explanatory variables, just as a multiple linear regression model does. However, their main disadvantage is that usually the estimation of these models is carried out by assuming a distribution for the duration, which in most cases is unknown.

An interesting new methodology is the one proposed by Stute [13], which can be used to estimate linear regression models with censored observations. It has good theoretical properties [13, 14] and it seems to be an interesting model to use in survival analysis. The model put forward by Stute can be considered as an AFT model but with the important characteristic that it allows us to estimate and make inference about the parameters of the model without assuming the distribution of the lifetime variable, usually unknown. Therefore, it avoids the problem of assuming a specific probability distribution, and, from this point of view, it could be considered an important alternative to the Cox PH models.

In addition, this method presents several advantages when compared to the PH model: (i) it does not need the assumption of proportional hazards; (ii) it models directly the effect of explanatory variables on the survival, so the interpretation of the results is clearer and easier (in terms of effects on mean survival time, as in the classical statistical models) than in the PH models, where we model the effect of covariates on a conditional probability. In addition, by using this methodology we could estimate the mean residual lifetime of a patient who has already survival up to time $t$; and (iii) it is simple to evaluate and it can be extended to consider more complex situations such as, for example, interactions between covariates and survival time, or to consider non-parametric effects of some covariates or covariates with time-dependent parameters.

Therefore, it can be of interest to compare, under several conditions, the performance of Stute's proposal with the ones based on PH and AFT parametric models. In order to do this, we compare these methods under situations where the PH hypothesis holds and where it does not hold, applying them to real examples and carrying out a simulation study. The rest of the paper is organized as follows. Section 2 describes two examples that motivate the new proposal. A brief review of the PH and AFT models is given and Stute's method is provided in Section 3. In Section 4, we fit different models to the data sets and comment on the

 

Table I. Survival times in months for breast cancer data.

| Negative staining | Positive staining | |
|---|---|---|
| 23 | 5 | 68 |
| 47 | 8 | 71 |
| 69 | 10 | 76+ |
| 70+ | 13 | 105+ |
| 71+ | 18 | 107+ |
| 100+ | 24 | 109+ |
| 101+ | 26 | 113 |
| 148 | 26 | 116+ |
| 181 | 31 | 118 |
| 198+ | 35 | 143 |
| 208+ | 40 | 154+ |
| 212+ | 41 | 162+ |
| 224+ | 48 | 188+ |
| | 50 | 212+ |
| | 59 | 217+ |
| | 61 | 225+ |

results obtained within the different specifications. In Section 5, we compare Cox's, AFT and Stute's methodologies under a simulation framework. Finally, in Section 6, we present some conclusions about the model and its advantages over the existing ones and some extensions are suggested.

## 2. BREAST AND GASTRIC CANCER DATA SETS

We present two examples that motivate the methodology proposed by Stute using well known data sets. In the first one, the essential assumption of the PH model is verified and in the second one this assumption does not hold.

### *2.1. Prognosis for women with breast cancer*

An investigation to evaluate a histochemical marker (the Helix pomatia agglutinin, HPA), which discriminates between primary breast cancer that either has metastasized or not, was carried out at the Middlesex Hospital [15]. The aim of this work was to investigate whether HPA staining can be used to predict the survival time of women who had breast cancer. In order to do this, women who had received a simple or radical mastectomy to treat a tumour between January 1969 and December 1971 were analysed. Sections of the tumours were treated with HPA and each tumour was subsequently classified as being positively or negatively stained. Positive staining corresponded to a tumour with the potential for metastasis. The study concluded in July 1987. Table I gives the survival times (in months) from surgery for each woman according to whether their tumour was positively or negatively stained. Censored survival times are labelled with the '+' symbol.

We are interested in whether or not there is a significant difference in the survival for the two groups of women. Thus, we need a model which tries to explain the effect of staining of the tumour on the survival time.

 

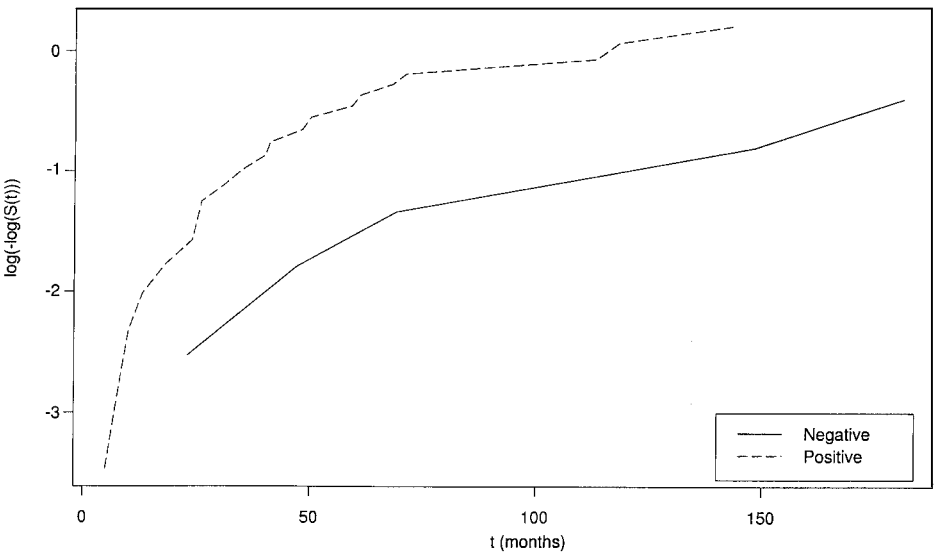


Figure 1. Graphical test of proportional hazard functions for breast cancer data.

As we do not know the distribution of the survival time variable, a possible model is the standard PH model; the fitting of this model assumes the proportionality of hazards for the two groups. In order to verify this assumption, we can use one of the different methods or tests proposed in the literature [6–11]. We have decided to use a simple graphical method [8] that plots the logarithm of the estimated cumulative hazard function for each group against the survival time. Parallel functions would mean that the assumption holds. Figure 1 indicates that PH model described above can be appropriate for this data.

### *2.2. Survival for gastric cancer patients*

In this example, treatments of locally advanced non-resectable gastric carcinoma consisting of chemotherapy alone versus a combination of chemotherapy and radiation are compared. The data (that is, the survival time in days for each treatment) are taken from a clinical trial reported by Stablein *et al.* [16] and are given in Table II.

We are interested in comparing the effect of both treatments on the survival of a patient. In this case, the distribution is also unknown and we may decide to use the PH model. However, before fitting a PH model, we check for the validity of the proportionality assumption using, as before, the graphic of the logarithm of the cumulative hazard function for each group versus the duration. Figure 2 clearly shows that these functions are non-parallel and that they even cross each other at some points. Therefore, we cannot assume PH in this example. This data set has been used by different authors [5, 9, 10, 16, 17] who, using different tests, have arrived at the same conclusion.

Hence, we need a model that allows for non-proportional hazards and that does not require the knowledge of the distribution of the survival time. Stute's model provides a possible solution for this situation and, in the next two sections, we proceed to describe this methodology and to compare its setting with the previous existing models.

 

Table II. Survival times in days for gastric cancer data.

| Combination | | Chemotherapy | |
|---|---|---|---|
| 17 | 307 | 1 | 499 |
| 42 | 315 | 63 | 524 |
| 44 | 401 | 105 | 529+ |
| 48 | 445 | 125 | 535 |
| 60 | 464 | 182 | 562 |
| 72 | 484 | 216 | 675 |
| 74 | 528 | 250 | 676 |
| 95 | 542 | 262 | 748 |
| 103 | 567 | 301 | 748 |
| 108 | 577 | 301 | 778 |
| 122 | 580 | 342 | 786 |
| 144 | 795 | 354 | 797 |
| 167 | 855 | 356 | 945+ |
| 170 | 882+ | 358 | 955 |
| 183 | 892+ | 380 | 958 |
| 185 | 1031+ | 381+ | 1180+ |
| 193 | 1033+ | 383 | 1245 |
| 195 | 1306+ | 383 | 1271 |
| 197 | 1335+ | 388 | 1277+ |
| 208 | 1366 | 394 | 1397+ |
| 234 | 1452+ | 408 | 1512+ |
| 235 | 1472+ | 460 | 1519+ |
| 254 | | 489 | |

## 3. GENERAL MODELS

For the standard version of the PH model, we have that the hazard function $\lambda(t)$ (the function which defines the probability that an individual dies at time $t$ assuming that he/she has survived up to that point in time) is defined as

$$\lambda(t, \mathbf{x}_i) = \lambda_0(t) \exp\left( \sum_{j=1}^{p} \beta_j x_{ij} \right) \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and represents the hazard function for an individual with covariate values all equal to zero, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ are the values measured on subject $i$ for the explanatory variables $\mathbf{X}_j$ $(j = 1, \ldots, p)$, and $\beta_j$ $(j = 1, \ldots, p)$ are unknown parameters in the model. The estimation of this model is carried out maximizing the partial likelihood [18].

In the AFT model, the hazard function is defined as

$$\lambda(t, \mathbf{x}_i) = \lambda_0\left( t \exp\left( \sum_{j=1}^{p} \beta_j x_{ij} \right) \right) \exp\left( \sum_{j=1}^{p} \beta_j x_{ij} \right) \tag{2}$$

where $\lambda_0(.)$ is a function of $t$, $X$, and $\beta$.

 

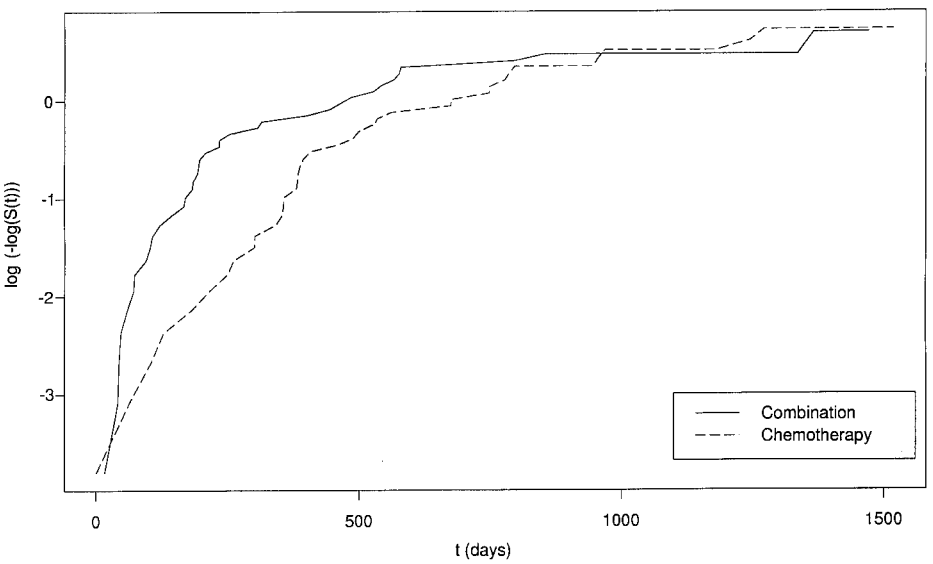


Figure 2. Graphical test of proportional hazard functions for gastric cancer data.

Thus, under this model, the effect of the explanatory variables on the survival time is direct, accelerating or decelerating the time to death or failure. Moreover, this model can be specified relating the logarithm of the survival time to its explanatory variables, just as a multiple linear regression model does. That is

$$\ln T = X\gamma + \varepsilon \tag{3}$$

where $X=[\mathbf{X}_1,\ldots,\mathbf{X}_p]$, $\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_p)'$ and $\gamma_j=-\beta_j$ for $j=1,\ldots,p$. In most situations, the estimation of this model is carried out by assuming a distribution for the duration and maximizing the log-likelihood.

The most commonly used parametric regression models in survival analysis (that is, the exponential, Weibull, log-normal, gamma or log-logistic models) can be considered as AFT models. In addition, the exponential and Weibull regression models can be considered as particular cases of both the AFT and PH models.

Unfortunately, because of the censoring effect, the actual lifetime $T$ is not always observable and instead we observe

$$Y_i=\min(T_i,C_i), \quad \delta_i=\begin{cases}1; & \text{if } T_i\leqslant C_i\\ 0; & \text{if } T_i>C_i\end{cases}$$

where $C_1,\ldots,C_n$ are the values of the censoring variable $C$, which is assumed independent of the duration variable $T$, and $\delta_i$ is an indicator of whether $T_i$ has been observed or not.

Within the AFT models' framework, Stute [13] presents a new methodology which requires very general hypotheses and where the estimators can be obtained using weighted least squares, that is, we use model (3) under the assumption that $E[\varepsilon|X]=0$. Here, the relation between the covariates and the duration, or some monotonic transformation of this, such as, for example,

 

the logarithmic one, is considered linear. Under this model, the estimator of $\gamma$ minimizes

$$\sum_{i=1}^{n} W_{in}[\ln Y_{(i)} - \mathbf{X}_{[i]}\gamma]^2 \tag{4}$$

where $\ln Y_{(i)}$ is the $i$th ordered value of the observed response variable $\ln Y$, $\mathbf{X}_{[i]}$ is the covariable associated with $\ln Y_{(i)}$ and $W_{in}$ are the Kaplan–Meier weights. These weights can be calculated using the expression

$$W_{in} = \hat{F}_n(\ln Y_{(i)}) - \hat{F}_n(\ln Y_{(i-1)}) = \frac{\delta_{[i]}}{n-i+1} \prod_{j=1}^{i-1} \left[\frac{n-j}{n-j+1}\right]^{\delta_{[j]}} \tag{5}$$

where $\hat{F}_n$ is a Kaplan–Meier estimator [19] of the distribution function $F$ for the variable $T$ and $\delta_{[i]}$ is the $\delta$ value associated with $\ln Y_{(i)}$. These weights can be also calculated using the redistribute to the right algorithm presented by Efron [20]. In this way, after calculating the $W_{in}$ weights, the minimization of (4) leads to the estimator of $\gamma$ given by

$$\hat{\gamma} = (X'WX)^{-1}X'W \ln Y$$

where $\ln Y = (\ln Y_{(1)}, \ldots, \ln Y_{(n)})'$, $W$ is a diagonal matrix with the Kaplan–Meier weights on its main diagonal and $X$ is defined as before. Stute studies the consistency of this estimator [13] and its asymptotic normal distribution [14]. As the asymptotic variance has a complicated expression to calculate, Stute [21] proposes the use of a simpler jack-knife estimator.

## 4. ANALYSIS OF THE DATA SETS

We now proceed to the analysis of the two data sets presented in Section 2.

### *4.1. Prognosis for women with breast cancer*

Under the PH model (1), the hazard of death at time $t$ for the $i$th woman is $\lambda(t, x_i) = \lambda_0(t)e^{x_i\beta}$, being $x_i = 0$ for negative staining and $x_i = 1$ for positive staining. That is, $\lambda_0(t)$ is the hazard function for a woman with a negatively stained tumour. If we fit this model, we obtain that the value of $\beta$, which maximizes the partial likelihood function [18], is $\hat{\beta} = 0.909$ and the estimate of its standard error is 0.501. Therefore, we can conclude that a woman who has a positively stained tumour will have a greater risk of death at any given time than a comparable woman whose tumour is negatively stained.

In addition, if we plot the logarithm of the cumulative hazard function against the logarithm of the survival time, we obtain a path close to a linear function. Therefore, this suggests the possible validity of the parametric Weibull regression model. Then, we can rewrite the model as a log-linear model, $\ln T = X\gamma + \sigma\varepsilon$, where $\varepsilon$ has a minimum extreme value distribution and $\sigma$ is a constant scale parameter. If we fit the Weibull regression model, we obtain the estimates $\hat{\gamma} = -0.9967$ (the negative sign is due to the fact that the estimator is now measuring the effect over the logarithm of the survival time) and $\hat{\sigma} = 1.0667$, with standard errors equal to 0.544 and 0.167, respectively. As an illustration, we have also fitted two alternative AFT models, the log-normal and log-logistic regression models, and have obtained $\hat{\gamma} = -1.151$ and $\hat{\gamma} = -1.149$, with standard errors 0.520 and 0.520, respectively.

 

We now use the approach proposed by Stute [13] to analyse this data set. That is, we consider the model (3), but we do not need to specify the distribution of $T$. If we minimize the sum of the weighted least squares to estimate $\gamma$ (see equation (4)), we obtain a differential effect (over the logarithm of the survival time) between groups of $\hat{\gamma}=-0.84$ with an estimated standard error of 0.407 (the standard error has been calculated using the jack-knife estimator [21]). Note that Stute's estimate shows: (i) a slightly smaller differential effect in the risk of death for both types of tumours and a smaller standard error when compared to the other approaches; and (ii) closer similarities to the AFT Weibull regression model, when compared to the AFT models. However, we obtain similar conclusions and results (positive staining indicates a poorer prognosis for breast cancer patients) using either one of these approaches.

### *4.2. Survival for gastric cancer patients*

We now compare the effect of both treatments and, thus, we define the covariate $X$ taking the value 0 if individuals are only treated with chemotherapy and 1 if they were treated with chemotherapy and radiation.

In order to illustrate the consequence of assuming PH when this does not hold, we fit the Cox standard PH model obtaining an estimate of $\hat{\beta}=0.266$ for the parameter which measures the differential effect between treatments, with an estimated standard error equal to 0.233. Thus, this differential effect is not significant, which seems a strange result if we look at the survival times in Table II. The lack of a significant effect of the combined treatments, when estimated using a PH model, is likely to be due to the lack of proportionality of the hazards in the two treatment groups. Thus, the analysis for treatment differences would be not significant as a result of the cancellation of an early advantage by a later disadvantage [16].

To solve this problem, Stablein *et al.* [16] proposed a non-proportional model for this data set. In particular, they proposed the use of the following model with a time varying hazards ratio by introducing some time-by-treatment interaction terms:

$$\lambda(t,x)=\lambda_0(t)\mathrm{e}^{\beta_1 x+\beta_2 x\frac{t}{30}+\beta_3 x(\frac{t}{30})^2} \tag{6}$$

Although the linear term in time would be sufficient to allow for a time varying relationship, the quadratic term allows the modelling of a relative hazard surface that rises and falls, that is, they consider a non-monotonic time dependence.

Estimation of the parameters will require the maximization of the partial likelihood function [22] giving the following estimates and standard errors:

$$\begin{aligned}
\hat{\beta}_1 &= 1.8664 \qquad & \hat{\sigma}_{\hat{\beta}_1} &= 0.6483\\
\hat{\beta}_2 &= -0.1767 \qquad & \hat{\sigma}_{\hat{\beta}_2} &= 0.0782\\
\hat{\beta}_3 &= 0.0028 \qquad & \hat{\sigma}_{\hat{\beta}_3} &= 0.0019
\end{aligned}$$

In this example, the additional flexibility for the hazard function indicates that there is a significant treatment difference. However, this effect disappears when the survival time of the patients increases, reaching slightly better results for long-term survivors by using the combined treatment [16].

 

As an alternative solution to the problem at hand, we have also fitted two AFT models that do not assume proportional hazards: the log-normal and log-logistic regression models, obtaining the estimated differential effect between treatments $\hat{\gamma}=-0.460$ and $\hat{\gamma}=-0.562$, with standard errors 0.2816 and 0.2451, respectively. The log-logistic model finds a significant differential effect, while the log-normal model does not.

It is also possible to introduce time-by-treatment interaction terms when fitting these AFT models; that is, we have considered the model

$$\ln t=\alpha+\gamma_1 x+\gamma_2 x\left(\frac{t}{30}\right)+\gamma_3 x\left(\frac{t}{30}\right)^2+\varepsilon \tag{7}$$

and, thus, obtained the following estimates of the parameters and their standard errors. For the log-normal AFT model:

$$\begin{aligned}
\hat{\alpha} &= 6.1631 & \hat{\sigma}_{\hat{\alpha}} &= 0.1442\\
\hat{\gamma}_1 &= -2.2402 & \hat{\sigma}_{\hat{\gamma}_1} &= 0.3349\\
\hat{\gamma}_2 &= 0.1758 & \hat{\sigma}_{\hat{\gamma}_2} &= 0.0403\\
\hat{\gamma}_3 &= -0.0018 & \hat{\sigma}_{\hat{\gamma}_3} &= 0.0009
\end{aligned}$$

and for the log-logistic AFT model:

$$\begin{aligned}
\hat{\alpha} &= 6.2472 & \hat{\sigma}_{\hat{\alpha}} &= 0.1035\\
\hat{\gamma}_1 &= -2.2656 & \hat{\sigma}_{\hat{\gamma}_1} &= 0.2163\\
\hat{\gamma}_2 &= 0.1726 & \hat{\sigma}_{\hat{\gamma}_2} &= 0.0250\\
\hat{\gamma}_3 &= -0.0020 & \hat{\sigma}_{\hat{\gamma}_3} &= 0.0006
\end{aligned}$$

From these results, the differential effect appears to be significant and, as in Stablein *et al.* [16], this effect vanishes as the survival time increases.

In order to avoid specific distributional assumptions, an alternative approach can be to apply Stute's methodology as described in the previous section. When we estimate the basic AFT model (that is, the one without interaction terms) the estimated differential effect between the two treatments is $\hat{\gamma}=-0.54$, and the estimated standard error (calculated, using the jack-knife estimator [21], as in the previous example) takes the value 0.27. Thus, we find a significant differential effect, where the treatment based on only chemotherapy obtains better results in terms of mean lifetime. This result, together with the ones obtained for the parametric AFT models, seems to indicate that the log-logistic regression model might be an adequate choice for this data set.

Moreover, we estimate the time-by-treatment interaction terms by considering model (7), and obtain the following estimates of the parameters and their standard errors:

$$\begin{aligned}
\hat{\alpha} &= 5.9417 & \hat{\sigma}_{\hat{\alpha}} &= 0.1918\\
\hat{\gamma}_1 &= -2.0402 & \hat{\sigma}_{\hat{\gamma}_1} &= 0.2392\\
\hat{\gamma}_2 &= 0.1871 & \hat{\sigma}_{\hat{\gamma}_2} &= 0.0165\\
\hat{\gamma}_3 &= -0.0025 & \hat{\sigma}_{\hat{\gamma}_3} &= 0.0003
\end{aligned}$$

 

These estimates show again a significant differential effect, where the treatment with only chemotherapy is significantly better. However, with the inclusion of interaction terms, we can, as in previous analysis, observe that this effect disappears when the survival time of the patients increases, reaching better results for long-term survivors by using the combined treatment. In summary, Stute's method leads to similar results as the other AFT models that use time-by-treatment interactions. However, it does not require the assumption of any specific distribution for the survival time.

The first example illustrates that, under the proportionality assumption, the results of the PH, Weibull AFT and Stute's methodologies are similar. However, when the proportionality is not verified by the data, fitting a PH model can lead to the wrong conclusions, as can be seen in the second example. For this case, the estimator minimizing (4) has good properties and, thus, provides a better approach. In addition, under this structure, we can also allow for a time-by-covariate interaction relationship by fitting a model like the one in (7).

One aspect that must be taken into account is that the results under the different methodologies are not directly comparable because the PH model explains the effect of covariates on the hazard function, while the AFT models measure the direct effect of the covariates on the duration. However, regardless of the methodology used, if both models are correct the inferential analysis should agree in terms of the significance of the covariates.

In addition, when using the AFT approach, Stute's methodology has advantages over the parametric ones because the latter are affected by a wrong specification of the probability distribution of the duration.

In order to analyse these issues, we have conducted a simulation study that is presented in the next section.

## 5. A SIMULATION STUDY

The objective of this section is twofold. First, we want to compare Stute's and Cox's methodologies in a context where the assumptions of both models hold and their estimators are directly comparable. Hence, this first purpose is to compare two methods that do not assume any probability distribution for the survival time. The second aim is to check the performance of Stute's method when the distribution of the survival time is known, that is, to compare Stute's estimators with those coming from a parametric AFT model.

For the first setting, note that the standard Cox specification belongs to the class of the PH models, while Stute's model can be included in the class of AFT models. However, since the exponential parametric regression model belongs to the two classes of models, we have decided to use this context to check the accuracy of both methodologies, so that we can compare their performance in a fair manner. In order to do that, we have generated an exponential regression model and the different estimates of the parameters will be compared under different settings.

Moreover, we would like to mention that, for this regression model, the parameter estimates obtained in both models are directly comparable, in absolute values, even though under the PH model specification (1) the estimation of $\beta$ measures the effect of the covariates on the hazard function. In addition, for the particular case of exponential distribution, the PH model can be rewritten as a log-linear model where $\hat{\gamma} = -\hat{\beta}$ captures the effect of the covariate on $\ln T$, as in (3).

 

Thus, the proposed model, in its log-linear specification, to generate the survival times is

$$\ln T = \gamma_1 X_1 + \gamma_2 X_2 + \varepsilon \tag{8}$$

where $X_1 \in U[0,2]$, $X_2 \in U[3,9]$, $\gamma_1 = 1$, $\gamma_2 = 3$ and $\varepsilon$ has a minimum extreme value distribution. In addition, we centre the error $\varepsilon$ to guarantee the assumption of $E[\varepsilon] = 0$, required under the Stute methodology. Moreover, we have to point out that, under this model, the baseline hazard function for the Cox model specification is constant and takes the value one.

The censoring variable is generated using a uniform distribution and as an independent variable from the survival time variable. The interval of uniformity depends on the censorship level in each case.

We would like to make a comparison between these two methodologies and study the effect of the censoring level and sample size on these differences. Therefore, the comparison is carried out under three different censorship levels (that is, 10, 30 and 45 per cent) and using three different sample sizes ($n = 50$, 100 and 200). We generated 2000 samples for each combination of censoring level and sample size.

The results are presented in Table III and Figures 3 to 5 (we only present the figures corresponding to the 30 per cent censorship level). Note that, in Table III(d), MSE denotes the sample mean square error. Each figure provides the results of the estimates for one level of censorship and sample size. The left panel presents the results for Stute's method and the right panel the corresponding one for Cox's method. For the purpose of illustration, we have also included in Table III the results for the AFT exponential regression model and, as expected, the latter provides better results.

The conclusions we can see from the results obtained from Stute's and Cox's methodologies are as follows: (i) when we increase the censoring level, the variance of the estimators tends to increase and the estimates lose precision; (ii) the estimation becomes better as the sample size increases; (iii) in terms of MSE, Stute's methodology provides more precise results for all cases considered (see Table III(d)); (iv) the differences, in terms of bias and standard errors, are larger for cases of explanatory variables with greater variability (see the estimates of $\beta_2$ and $\gamma_2$ in Tables III(a)–(c)). We have replicated this simulation study with normally distributed covariates and have obtained very similar conclusions.

For the second purpose, we have simulated data from the model

$$\ln T = \gamma_1 X_1 + \gamma_2 X_2 + \sigma\varepsilon \tag{9}$$

where $X_1 \in U[0,2]$, $X_2 \in U[3,9]$, $\gamma_1 = 1$, $\gamma_2 = 3$ and the distribution of $\varepsilon$ will be fitted according to a log-normal and a log-logistic distribution. For the log-normal case, $\varepsilon$ follows a standard normal distribution and we have chosen $\sigma$ equal to 1. For the log-logistic case, $\varepsilon$ follows a logistic distribution and the value for $\sigma$ is chosen to be equal to 0.5.

The results are presented in Table IV (for the log-normal case) and in Table V (for the log-logistic case).

As expected, the results are better when we estimate the parametric model using the information of the correct probability distribution that generated the data. However, the accuracy of the results coming from Stute's method justifies its validity and suggest its use when the probability distribution is unknown.

 

Table III. Estimates of the parameters in equation (8) under uniform covariates ($X_1 \in U[0,2]$, $X_2 \in U[3,9]$) and an exponential distribution assumption, using three alternative estimation procedures: $\gamma_1 = 1$, $\gamma_2 = 3$.

*(a) Estimates with sample size $n = 50$*

| Censorship | | Cox | | Stute | | Exponential | |
|---|---|---|---|---|---|---|---|
| | | $\beta_1 = -\gamma_1$ | $\beta_2 = -\gamma_2$ | $\gamma_1$ | $\gamma_2$ | $\gamma_1$ | $\gamma_2$ |
| 10% | Bias | 0.0263 | 0.1284 | −0.0075 | −0.0012 | −0.0121 | 0.0037 |
| | Standard error | 0.3737 | 0.4867 | 0.3492 | 0.1390 | 0.2789 | 0.1119 |
| 30% | Bias | 0.0465 | 0.1821 | −0.0157 | −0.0018 | −0.0199 | −0.0023 |
| | Standard error | 0.4535 | 0.5900 | 0.4075 | 0.1653 | 0.3352 | 0.1718 |
| 45% | Bias | 0.0885 | 0.2904 | −0.0297 | −0.0373 | −0.0438 | −0.0503 |
| | Standard error | 0.5823 | 0.7814 | 0.4873 | 0.2202 | 0.4383 | 0.3541 |

*(b) Estimates with sample size $n = 100$*

| Censorship | | Cox | | Stute | | Exponential | |
|---|---|---|---|---|---|---|---|
| | | $\beta_1 = -\gamma_1$ | $\beta_2 = -\gamma_2$ | $\gamma_1$ | $\gamma_2$ | $\gamma_1$ | $\gamma_2$ |
| 10% | Bias | 0.0153 | 0.0472 | −0.0072 | 0.0002 | −0.0006 | −0.0009 |
| | Standard error | 0.2461 | 0.2953 | 0.2647 | 0.0752 | 0.2082 | 0.0594 |
| 30% | Bias | 0.0228 | 0.0648 | −0.0113 | −0.0003 | 0.0019 | −0.0023 |
| | Standard error | 0.2909 | 0.3399 | 0.3060 | 0.0919 | 0.2425 | 0.0805 |
| 45% | Bias | 0.0224 | 0.0795 | −0.0281 | −0.0247 | −0.0032 | −0.0281 |
| | Standard error | 0.3219 | 0.3777 | 0.3794 | 0.1245 | 0.2812 | 0.2231 |

*(c) Estimates with sample size $n = 200$*

| Censorship | | Cox | | Stute | | Exponential | |
|---|---|---|---|---|---|---|---|
| | | $\beta_1 = -\gamma_1$ | $\beta_2 = -\gamma_2$ | $\gamma_1$ | $\gamma_2$ | $\gamma_1$ | $\gamma_2$ |
| 10% | Bias | 0.0089 | 0.0268 | −0.0006 | 0.0001 | −0.0001 | 0.0003 |
| | Standard error | 0.1510 | 0.1985 | 0.1596 | 0.0556 | 0.1276 | 0.0445 |
| 30% | Bias | 0.0105 | 0.0315 | −0.0021 | −0.0002 | −0.0008 | −0.00009 |
| | Standard error | 0.1720 | 0.2244 | 0.1876 | 0.0648 | 0.1454 | 0.0520 |
| 45% | Bias | 0.0140 | 0.0425 | −0.0133 | −0.0141 | −0.0032 | −0.0136 |
| | Standard error | 0.1906 | 0.2460 | 0.2254 | 0.0898 | 0.1665 | 0.1519 |

*(d) Comparison of the MSE for Cox, Stute and the exponential AFT methods.*

| Sample | Censorship 10 per cent | | | Censorship 30 per cent | | | Censorship 45 per cent | | |
|---|---|---|---|---|---|---|---|---|---|
| size | Cox | Stute | Exp | Cox | Stute | Exp | Cox | Stute | Exp |
| $n = 50$ | 0.3937 | 0.1413 | 0.0900 | 0.5891 | 0.1936 | 0.1422 | 1.0417 | 0.2882 | 0.3219 |
| $n = 100$ | 0.1502 | 0.0757 | 0.0468 | 0.2049 | 0.1022 | 0.0652 | 0.2532 | 0.1608 | 0.1296 |
| $n = 200$ | 0.0629 | 0.0285 | 0.0183 | 0.0810 | 0.0393 | 0.0238 | 0.0988 | 0.0592 | 0.0509 |

 

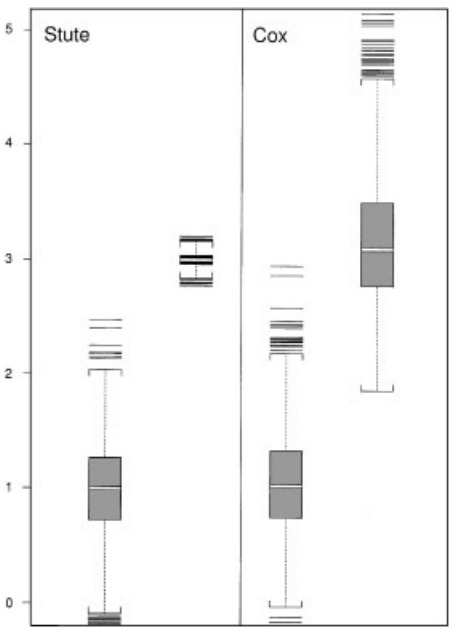


Figure 3. Box-plots of the estimations of the regression coefficients for the model $\ln T = \gamma_1 X_1 + \gamma_2 X_2 + \varepsilon$ for a sample size $n = 50$ and a censoring level $= 30\%$. The left panel presents the results for Stute's method and the right panel for Cox's method. Within each panel, the left plot contains the estimates for $\gamma_1$ and the right one for $\gamma_2$.

## 6. CONCLUSIONS

Cox's PH regression model is the most common way of analysing prognostic factors in clinical research. This is probably due to the fact that this model allows us to estimate and make inference about the parameters without assuming any distribution for the lifetime, whose distribution is often unknown. However, it does have the requirement of proportional hazards, which is not always satisfied by the data. In these situations, AFT models provide an alternative framework to fit the data. Moreover, under these models we measure the direct effect of the explanatory variables on the survival time and not on a conditional probability, as we do in the Cox PH model. This characteristic allows for an easier interpretation of the results because the parameters measure the effect of the correspondent covariate on the mean lifetime.

Among the AFT models, log-normal and log-logistic regression models are very common choices, when the proportional hazards assumption does not hold but they assume specific distributions for the duration, which is unknown in many cases. In order to avoid this disadvantage, Stute [13] proposed a method that allows us to estimate and make inference about the parameters without assuming any distribution for the survival time.

In this paper we have analysed the performance of these alternative methodologies in two different frameworks. As a first approach, two clinical data sets have been analysed. In the

 

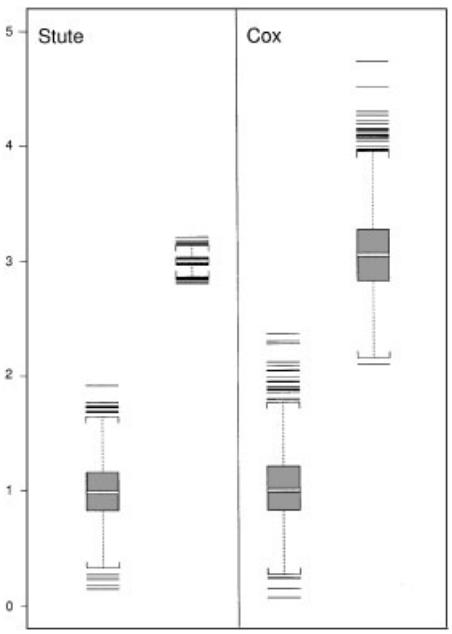


Figure 4. Box-plots of the estimations of the regression coefficients for the model $\ln T = \gamma_1 X_1 + \gamma_2 X_2 + \varepsilon$ for a sample size $n = 100$ and a censoring level $= 30\%$. The left panel presents the results for Stute's method and the right panel for Cox's method. Within each panel, the left plot contains the estimates for $\gamma_1$ and the right one for $\gamma_2$.

first example, the hypotheses of Cox's model hold and Stute's approach performed better in terms of standard errors. The second data set is an example where the PH assumption does not hold and, after applying the standard Cox method, the results seem to lead to the wrong conclusions while Stute's method leads to the correct conclusions providing more precise estimators.

Finally, a simulation study, where the estimates are comparable, shows that Stute's method can be successfully applied in a context where Cox assumptions also hold. Moreover, we can observe that the methodology proposed by Stute is more precise in all cases considered than the one proposed by Cox. For the cases where the PH assumption does not hold, Stute's method is compared with the parametric AFT under the correct assumption of the probability distribution. In this unfair context, and as expected, Stute's estimates are less precise. However, this loss of precision is very small, which, in a sense, advocates the use of this methodology when the probability distribution is unknown, the most usual situation in practice. Thus, we can avoid the problem of a wrong specification of the model when assuming an incorrect distribution for the survival time. All the results presented above indicate that this method provides good estimates under very general and different situations.

In addition, as in the standard Cox PH, this model can also be modified in a natural way to consider more complex situations such as, for example, non-linear dependence between

 

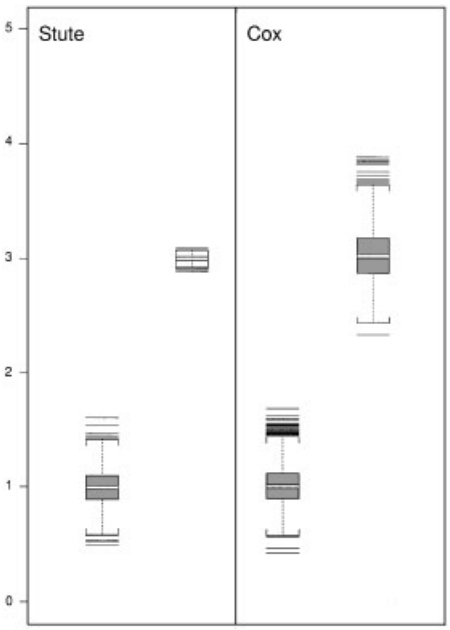


Figure 5. Box-plots of the estimations of the regression coefficients for the model $\ln T = \gamma_1 X_1 + \gamma_2 X_2 + \varepsilon$ for a sample size $n = 200$ and a censoring level $= 30\%$. The left panel presents the results for Stute's method and the right panel for Cox's method. Within each panel, the left plot contains the estimates for $\gamma_1$ and the right one for $\gamma_2$.

the lifetime and the explanatory variables [23], where $f(.)$ is a known non-linear function. It can be also extended to a partial model by adding a non-parametric component [24]. Thus, by using this additional term, we can consider situations where the functional effect of some covariate is unknown or it is restrictive to assume some specific functional form, $\ln T = X\gamma + f(Z) + \varepsilon$, situations with time-depending parameters for some covariate, $\ln T = X\gamma + Z\eta(t) + \varepsilon$; or situations where the intercept changes with time in a non-parametric way, $\ln T = f(t) + X\gamma + \varepsilon$.

Hence the main conclusion is that Stute's methodology appears as a very useful tool to be taken into account in survival analysis and that some appealing extensions are possible, although further research must be carried out along these lines.

ACKNOWLEDGEMENTS

This work was partially supported by Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), Dirección General de Enseñanza Superior e Investigación Científica del Ministerio Español de Educación y Cultura and Gobierno Vasco under research grants UPV 038.321-HA129/99 and UPV 038.321-13631/2001, PB98-0149, PI-1999-46 and PI-1999-70. The authors thank two anonymous referees and the editor for providing thoughtful comments and suggestions which led to substantial improvement of the presentation of the material in this paper.

 

Table IV. Estimates of the parameters in equation (9) under uniform covariates ($X_1 \in U[0,2]$, $X_2 \in U[3,9]$) and a log-normal distribution assumption, using two alternative estimation procedures: $\gamma_1 = 1$, $\gamma_2 = 3$.

*(a) Estimates with sample size $n = 50$*

| Censorship | | Stute | | Log-normal | |
|---|---|---|---|---|---|
| | | $\gamma_1$ | $\gamma_2$ | $\gamma_1$ | $\gamma_2$ |
| 10% | Bias | 0.0019 | −0.0030 | 0.0011 | −0.0032 |
| | Standard error | 0.2547 | 0.0826 | 0.2527 | 0.0817 |
| 30% | Bias | −0.0048 | −0.0055 | −0.0053 | −0.0027 |
| | Standard error | 0.2982 | 0.1034 | 0.2785 | 0.0962 |
| 45% | Bias | 0.0018 | −0.0270 | −0.0035 | −0.0002 |
| | Standard error | 0.3563 | 0.1434 | 0.3092 | 0.1227 |

*(b) Estimates with sample size $n = 100$*

| Censorship | | Stute | | Log-normal | |
|---|---|---|---|---|---|
| | | $\gamma_1$ | $\gamma_2$ | $\gamma_1$ | $\gamma_2$ |
| 10% | Bias | 0.0028 | −0.0009 | 0.0036 | −0.0011 |
| | Standard error | 0.1823 | 0.0608 | 0.1809 | 0.0606 |
| 30% | Bias | −0.0002 | −0.0034 | 0.0008 | −0.0017 |
| | Standard error | 0.2175 | 0.0728 | 0.2044 | 0.0696 |
| 45% | Bias | 0.0041 | −0.0215 | 0.0031 | −0.0009 |
| | Standard error | 0.2671 | 0.0991 | 0.2246 | 0.0826 |

*(c) Estimates with sample size $n = 200$*

| Censorship | | Stute | | Log-normal | |
|---|---|---|---|---|---|
| | | $\gamma_1$ | $\gamma_2$ | $\gamma_1$ | $\gamma_2$ |
| 10% | Bias | −0.0014 | 0.0006 | −0.0017 | 0.0004 |
| | Standard error | 0.1377 | 0.0436 | 0.1367 | 0.0436 |
| 30% | Bias | −0.0023 | −0.0019 | −0.0017 | −0.0011 |
| | Standard error | 0.1653 | 0.0527 | 0.1566 | 0.0501 |
| 45% | Bias | −0.0131 | −0.0174 | −0.0044 | 0.0003 |
| | Standard error | 0.2109 | 0.0736 | 0.1733 | 0.0603 |

*(d) Comparison of the MSE for Stute and the log-normal AFT methods.*

| Sample size | Censorship 10 per cent | | Censorship 30 per cent | | Censorship 45 per cent | |
|---|---|---|---|---|---|---|
| | Stute | Log-normal | Stute | Log-normal | Stute | Log-normal |
| $n = 50$ | 0.0717 | 0.0705 | 0.0996 | 0.0868 | 0.1482 | 0.1106 |
| $n = 100$ | 0.0369 | 0.0364 | 0.0526 | 0.0466 | 0.0816 | 0.0572 |
| $n = 200$ | 0.0208 | 0.0205 | 0.0301 | 0.0270 | 0.0503 | 0.0336 |

 

Table V. Estimates of the parameters in equation (9) under uniform covariates ($X_1 \in U[0,2]$, $X_2 \in U[3,9]$) and a log-logistic distribution assumption, using two alternative estimation procedures: $\gamma_1 = 1$, $\gamma_2 = 3$.

*(a) Estimates with sample size $n = 50$*

| Censorship | | Stute | | Log-logistic | |
|---|---|---|---|---|---|
| | | $\gamma_1$ | $\gamma_2$ | $\gamma_1$ | $\gamma_2$ |
| 10% | Bias | −0.0049 | −0.0010 | −0.0026 | −0.0008 |
| | Standard error | 0.2479 | 0.0990 | 0.2377 | 0.0950 |
| 30% | Bias | −0.0097 | −0.0011 | −0.0012 | −0.0004 |
| | Standard error | 0.2917 | 0.1199 | 0.2733 | 0.1100 |
| 45% | Bias | −0.0122 | −0.0193 | 0.0079 | 0.0025 |
| | Standard error | 0.3491 | 0.1557 | 0.3090 | 0.1309 |

*(b) Estimates with sample size $n = 100$*

| Censorship | | Stute | | Log-logistic | |
|---|---|---|---|---|---|
| | | $\gamma_1$ | $\gamma_2$ | $\gamma_1$ | $\gamma_2$ |
| 10% | Bias | −0.0044 | −0.0004 | −0.0048 | −0.0002 |
| | Standard error | 0.1877 | 0.0534 | 0.1785 | 0.0511 |
| 30% | Bias | −0.0054 | −0.0009 | −0.0068 | 0.0006 |
| | Standard error | 0.2241 | 0.0670 | 0.2026 | 0.0615 |
| 45% | Bias | −0.0185 | −0.0146 | −0.0064 | 0.0034 |
| | Standard error | 0.2764 | 0.0894 | 0.2233 | 0.0744 |

*(c) Estimates with sample size $n = 200$*

| Censorship | | Stute | | Log-logistic | |
|---|---|---|---|---|---|
| | | $\gamma_1$ | $\gamma_2$ | $\gamma_1$ | $\gamma_2$ |
| 10% | Bias | −0.0003 | 0.0003 | 0.0008 | 0.0001 |
| | Standard error | 0.1139 | 0.0399 | 0.1070 | 0.0379 |
| 30% | Bias | −0.0008 | −0.0006 | 0.00007 | 0.0004 |
| | Standard error | 0.1348 | 0.0475 | 0.1221 | 0.0438 |
| 45% | Bias | −0.0090 | −0.0106 | 0.0018 | 0.0013 |
| | Standard error | 0.1646 | 0.0670 | 0.1327 | 0.0529 |

*(d) Comparison of the MSE for Stute and the log-logistic AFT methods.*

| Sample | Censorship 10 per cent | | Censorship 30 per cent | | Censorship 45 per cent | |
|---|---|---|---|---|---|---|
| size | Stute | Log-logistic | Stute | Log-logistic | Stute | Log-logistic |
| $n = 50$ | 0.0708 | 0.0655 | 0.0995 | 0.0867 | 0.1466 | 0.1126 |
| $n = 100$ | 0.0381 | 0.0344 | 0.0547 | 0.0448 | 0.0849 | 0.0554 |
| $n = 200$ | 0.0145 | 0.0128 | 0.0204 | 0.0168 | 0.0317 | 0.0204 |

 

## REFERENCES


1. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society*, *Series B* 1972; **34**: 187–220.
2. Lawless JF. *Statistical Models and Methods for Lifetime Data Analysis*. Wiley: New York, 1982.
3. Altman DG, De Stavola BL, Love SB, Stepniewska KA. Review of survival analyses published in cancer journals. *British Journal of Cancer* 1985; **72**:511–518.
4. Abrahamowicz M, Mackenzie T, Esdaile JM. Time-dependent hazard ratio: modelling and hypothesis testing with application in Lupus Nephritis. *Journal of the American Statistical Association* 1996; **91**:1432–1439.
5. Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine* 1994; **13**:1045–1062.
6. Nagelgerke NJD, Oosting J, Hart AAM. A simple test for goodness of fit of Cox's proportional hazards model. *Biometrics* 1984; **40**:483–486.
7. Wei J. Testing goodness of fit for proportional hazards model with censored observations. *Journal of the American Statistical Association* 1984; **79**:649–652.
8. Kay R. Proportional hazard regression models and the analysis of censored survival data. *Applied Statistics* 1977; **26**:227–237.
9. Hess KR. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Statistics in Medicine* 1995; **14**:1707–1723.
10. Moreau T, O'Quigley J, Mesbah M. A global goodness-of-fit statistics for the proportional hazards model. *Applied Statistics* 1985; **3**:212–218.
11. Kooperberg C, Stone CJ, Truong YK. Hazard regression. *Journal of the American Statistical Association* 1995; **90**:78–94.
12. Gray RJ. Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association* 1992; **87**:942–951.
13. Stute W. Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis* 1993; **45**:89–103.
14. Stute W. Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics* 1996; **23**:461–471.
15. Leathem AJ, Brooks SA. Predictive value of lectin binding on breast-cancer recurrence and survival. *Lancet* 1987; **1**:1054–l056.
16. Stablein DM, Carter WH, Jr, Novak JW. Analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials* 1981; **2**:149–159.
17. Mantel N, Stablein DM. The crossing hazard function problem. *Statistician* 1988; **37**:59–64.
18. Cox DR. Partial likelihood. *Biometrika* 1975; **62**:269–276.
19. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**:457–481.
20. Efron B. The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1967; **4**:831–853.
21. Stute W. The jack-knife estimate of variance of a Kaplan–Meier integral. *Annals of Statistics* 1996; **24**:2679–2704.
22. Stablein DM, Carter WH, Jr, Wampler GL. Survival analysis of drug combinations using hazard model with time dependent covariates. *Biometrics* 1980; **36**:537–549.
23. Stute W. Nonlinear censored regression. *Statistica Sinica* 1999; **9**:1089–1102.
24. Orbe J. Un modelo de regresión parcial censurado para análisis de supervivencia. PhD dissertation, Universidad del País Vasco, Bilbao, Spain, 2000.


 