NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Residuals and Functional Form in Accelerated Life Regression Models

by

Bo Henry Lindqvist, Stein Aaserud and Jan Terje Kvaløy

PREPRINT
STATISTICS NO. 13/2012

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
http://www.math.ntnu.no/preprint/statistics/2012/S13-2012.pdf
Bo H. Lindqvist has homepage: http://www.math.ntnu.no/ bo
E-mail: bo@math.ntnu.no
Address: Department of Mathematical Sciences, NTNU, N-7491 Trondheim,
Norway.

Residuals and Functional Form in Accelerated Life Regression Models

Bo Henry Lindqvist
Department of Mathematical Sciences
Norwegian University of Science and Technology
N-7491 Trondheim, Norway
Email: bo@math.ntnu.no

Stein Aaserud*
Department of Mathematical Sciences
Norwegian University of Science and Technology
N-7491 Trondheim, Norway
Email: stein.aaserud@aibel.com

Jan Terje Kvaløy
Department of Mathematics and Natural Sciences
University of Stavanger
N-4036 Stavanger, Norway
Email: jan.t.kvaloy@uis.no

December 20, 2012

Abstract

We study residuals of parametric accelerated failure time (AFT) models for censored data, with the main aim of inferring the correct functional form of possibly misspecified covariates. We demonstrate the use of the methods by a simulated example and by applications to two reliability data sets. We also consider briefly a corresponding approach for parametric proportional hazards models.

Keywords: Cox-Snell residuals; exponential regression; misspecified model; proportional hazards model

*Current address: Aibel AS, P.O.Box 444, N-1373 Billingstad, Norway

1 Introduction

Accelerated failure time (AFT) models are commonly used for modelling a possible relationship between event times and covariates. Applications include a variety of areas, such as reliability engineering, biostatistics, economics and social sciences. The AFT failure time model can be written

$$\log Y = f(\boldsymbol{X}) + \sigma W, \tag{1}$$

where Y is the event time; $\boldsymbol{X} = (X_1, \ldots, X_p)$ is a vector of covariates; $f(\cdot)$ is some function determining the influence of the covariates; while σW is an "error" term. The parameter σ is here considered as a scale parameter, while W is assumed to have a fully specified "standardized" distribution, such as the standard normal distribution, in which case Y is lognormal; the standard Gumbel distribution for the smallest extreme (in the following called the Gumbel distribution), in which case Y is Weibull-distributed; and the standard logistic distribution, in which case Y is called log-logistically distributed. Our generic notation for the distribution function of W will be $\Phi(u) = P(W \leq u)$. For short we shall say that W has distribution Φ.

Although the methods we present in this paper will appear to be nonparametric in nature, our basic concern will be on fully parametric AFT models. This means that $f(\cdot)$ is basically assumed to be a parametric function, usually of the linear form

$$f(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \tag{2}$$

Nice introductions to parametric AFT model can be found in the monographs Collett (2003, Ch. 6); Meeker and Escobar (1998, Ch. 17).

For proper analysis of survival data it is of course important that a reasonably correct model is used. In this paper we will focus on methods for checking and suggesting the functional form for the covariates in the representation (2). By our approach, this may alternatively be viewed as a search for a "best possible" additive model of the form

$$\log Y = f_1(X_1) + \cdots + f_p(X_p) + \sigma W \tag{3}$$

for functions $f_j(\cdot)$, $j = 1, \ldots, p$, in the following called *covariate functions*.

Our first concern regards procedures for checking AFT models in the form of (1) and (2) by means of residual plots. There are in fact several kinds of residuals appropriate for such model checking, see for example Collett (2003, Ch. 7). The most natural residual is the so called standardized residual. This will play a role in the estimation of covariate functions, but we shall for a large part be concerned with the Cox-Snell residuals, also called generalized residuals, originally suggested by Cox and Snell (1968), and probably being the most widely used residuals in survival analysis.

The typical application of Cox-Snell residuals is to do model checking by deciding whether the full set of Cox-Snell residuals, possibly censored, deviates significantly from what would be expected if they were exponentially distributed (e.g. Kay, 1977). In the present paper we are more concerned with the alternative use of the residuals, namely to plot them versus single covariates. This can in fact be done in several ways, in particular because the data may include censored values. For continuous covariates we shall consider two basic methods for residual plotting based on smoothing the residuals as functions of the particular covariates.

We shall also briefly treat residual plots for discrete covariates. This is first of all of interest when the covariate takes only a relatively small finite number of values, but can also be used in connection with stratification of data with respect to covariates (e.g. Arjas, 1988).

The main reason for the interest in Cox-Snell residuals is that their ideal distribution is the exponential, whatever be the distribution of W. This makes it possible to use similar methods when models differ in the distribution of W. We shall also see how the special properties of the exponential distribution simplify and unify the handling of censoring. On the other hand, standardized residuals will ideally have distribution Φ and should hence be treated differently in different model types.

It should be mentioned that the behaviour of Cox-Snell residuals for checking overall goodness of fit of survival models in general has been critizised, particularly in the case of the semiparametric Cox-model (see e.g. Crowley and Storer, 1983). The possible problems are due to the nonparametric estimation of the baseline hazard function in Cox-models which leads to a violation of the approximate exponentiality of the Cox-Snell residuals. On the other hand, Crowley and Storer (1983) report on more satisfactory behaviour when residuals are plotted against covariate values. Many of these problems will be less pronounced for the parametric models considered here, due to a finite number of parameters in the baseline cases.

Besides the study of residual plots for censored AFT models, the main purpose of the paper is to show how the, possibly smoothed, residuals can be used to derive appropriate covariate functions $f_j(\cdot)$ in the representation (3). We hence seeks to complement results and methods for the semiparametric Cox-model as earlier presented in Therneau et al. (1990), Grambsch et al. (1995), and later described in the monograph Therneau et al. (2000, Ch. 5). For a general discussion of methods for goodness-of-fit in survival models based on residuals, we also refer to Andersen et al. (1993, VII.3.4).

The outline of the paper is as follows. In Section 2 we state a set of precise assumptions for the distribution of the random variables that go into our models, and we also derive the likelihood function for the observed data. In Section 3 we give the basic definitions and properties for residuals to be used for AFT models, and discuss how modifications are made for censored observations. Section 4 is concerned with plotting of residuals, both for continuous and discrete covariates. Particular emphasis is given to the constructiion of informative residual plots in cases with censored observations. In Section 5 we then show how residuals of possibly misspecified models can be used to infer the appropriate functional form of covariates in an AFT model. This can be viewed as the main section of the paper, where various methods are presented. Section 6 studies the special case when lifetimes are Weibull distributed and the results are applied to a real dataset as well as a simulated one. An adaptation of the approach of Section 5 to cover parametric proportional hazards models is considered in Section 7. This treatment complements similar studies for the semiparametric Cox model performed by Therneau et al. (1990) and Grambsch et al. (1995). A few concluding remarks are given in the final section, Section 8. Some additional results are given in Appendix A-D, in particular an introduction to the covariate order method for exponential regression.

2 Model assumptions

In order for a rigorous treatment, we shall in the following make precise assumptions on the probability mechanisms that produce our data. It should be noted that to a large extent the conditions are stated to simplify arguments and make a more transparent theory. Some of the conditions therefore appear stronger than needed. It is, however, beyond the scope of this paper to consider weakest possible assumptions.

The observation for an individual is assumed to be a realization of an underlying random vector $(\boldsymbol{X}, W, C)$. Here $\boldsymbol{X}$ (the covariate vector) can have both discrete and absolutely continuous components, and has a distribution given by a density $g_{\boldsymbol{X}}(\cdot)$ with respect to a product of counting measures and Lebesgue measures, according to the types of covariates. Further, W ("error") has an absolutely continuous distribution with distribution function $\Phi(\cdot)$ and density function $\phi(\cdot)$, where we make the assumption that $\phi(u) > 0$ for all $-\infty < u < \infty$. Finally, C (censoring time) is an absolutely continuous non-negative random variable, which may depend on $\boldsymbol{X}$, with conditional survival function given $\boldsymbol{X} = \boldsymbol{x}$ denoted $G_C(c|\boldsymbol{x})$. Furthermore, W is assumed to be independent of $(\boldsymbol{X}, C)$.

The true lifetime defined for the individual is Y with

$$\log Y = f(\boldsymbol{X}) + \sigma W, \tag{4}$$

for a given function f and a positive (scale) parameter σ. The observed lifetime for the individual is $T = \min(Y, C)$, while $\Delta = I(Y < C)$ is the status defined for this individual.

Under these assumptions it is straightforward to show that the joint density of the observable vector for the individual, $(T, \Delta, \boldsymbol{X})$, at $(t, \delta, \boldsymbol{x})$, is

$$g_{\boldsymbol{X}}(\boldsymbol{x}) \left\{g_Y(t|\boldsymbol{x})G_C(t|\boldsymbol{x})\right\}^{\delta} \left\{g_C(t|\boldsymbol{x})G_Y(t|\boldsymbol{x})\right\}^{1-\delta} \tag{5}$$

where lower case g means density, while capital G means survival function, for the respective random variable which is given as index. Note that we have used that T and C are independent given $\boldsymbol{X}$, which follows from the above assumptions.

It is clear that for an i.i.d. sample $\{(t_i, \delta_i, \boldsymbol{x}_i); i = 1, \ldots, n\}$ from this joint distribution, the likelihood function is given as

$$\prod_{i=1}^{n} g_{\boldsymbol{X}}(\boldsymbol{x}_i) \left\{g_Y(t_i|\boldsymbol{x}_i)G_C(t_i|\boldsymbol{x}_i)\right\}^{\delta_i} \left\{g_C(t_i|\boldsymbol{x}_i)G_Y(t_i|\boldsymbol{x}_i)\right\}^{1-\delta_i}.$$

This likelihood will be the basis for maximum likelihood estimation in the parametric regression models we shall encounter. However, since we shall assume that the functions $f_{\boldsymbol{X}}(\cdot)$, $F_C(\cdot|\cdot)$, $f_C(\cdot|\cdot)$ do not depend on the parameters of interest (which are of course the ones of g_Y and G_Y), the resulting likelihood used for maximization will be of the following well known form:

Under the standard assumption that the functions $g_{\boldsymbol{X}}(\cdot)$, and $g_C(\cdot|\cdot)$ do not depend on the parameters of $g_Y(\cdot|\cdot)$ we obtain the standard likelihood for survival analysis,

$$\prod_{i=1}^{n} \left\{g_Y(t_i|\boldsymbol{x}_i)\right\}^{\delta_i} \left\{G_Y(t_i|\boldsymbol{x}_i)\right\}^{1-\delta_i} \tag{6}$$

3 Residuals in AFT models

3.1 Standardized and Cox-Snell residuals

Standardized residuals in AFT models are based on solving equation (1) for W. It is clear that for an observed lifetime T from model (1), if we let

$$S = \frac{\log Y - f(\boldsymbol{X})}{\sigma}, \tag{7}$$

then, conditionally on $\boldsymbol{X}$, S has the distribution Φ.

Let data $(t_i, \delta_i, \boldsymbol{x}_i),\ i = 1, \ldots, n$, as considered in Section 2, be given. These are possibly right-censored, with the δ_i being censoring indicators. The standardized residuals are then defined by $(\hat{s}_i, \delta_i),\ i = 1, \ldots, n$, where

$$\hat{s}_i = \frac{\log t_i - \hat{f}(\boldsymbol{x}_i)}{\hat{\sigma}}, \tag{8}$$

with $\hat{f}(\cdot), \hat{\sigma}$ being appropriate estimators of the underlying f and σ, respectively. The idea is that if the model used for estimation is correctly specified, then the set $(\hat{s}_i, \delta_i),\ i = 1, \ldots, n$ should behave similar to a censored sample from the distribution Φ. Censoring for the $\hat{s}_i$ here corresponds to the fact that if a t_i is a right censored observation (i.e. $\delta_i = 0$), then $\hat{s}_i$ becomes "too small".

The Cox-Snell residuals are based on the fact that if Y is a lifetime, with corresponding survival function $G(t) = P(Y > t)$, then the random variable $-\log G(Y)$ is unit exponentially distributed, i.e. exponentially distributed with mean 1, whatever be $G(t)$.

The Cox-Snell residuals for the model (1) are hence obtained by first noting that

$$G(t|\boldsymbol{X}) \equiv P(Y > t|\boldsymbol{X}) = 1 - \Phi\left(\frac{\log t - f(\boldsymbol{X})}{\sigma}\right),$$

which hence implies that

$$R = -\log G(Y|\boldsymbol{X}) = -\log\left(1 - \Phi\left(\frac{\log Y - f(\boldsymbol{X})}{\sigma}\right)\right) \tag{9}$$

is, conditionally on $\boldsymbol{X}$, unit exponentially distributed.

For the data and fitted model as given above, the Cox-Snell residuals are therefore given as $(\hat{r}_i, \delta_i),\ i = 1, \ldots, n$, where

$$\hat{r}_i = -\log\left(1 - \Phi\left(\frac{\log t_i - \hat{f}(\boldsymbol{x_i})}{\hat{\sigma}}\right)\right). \tag{10}$$

If the model is correctly specified, then the set $(\hat{r}_i, \delta_i),\ i = 1, \ldots, n$ should behave similar to a censored sample of unit exponentially distributed variables.

Note that we have the following relations between the "theoretical" standardized residuals and Cox-Snell residuals,

$$R = -\log(1 - \Phi(S)) \tag{11}$$

$$S = \Phi^{-1}(1 - e^{-R}) \tag{12}$$

(and corresponding relations between the $\hat{r}_i$ and $\hat{s}_i$). It is shown in Appendix A that under a certain condition on Φ (which is valid for the lognormal, Weibull and log-logistic cases), R given in (11) is a strictly convex function of S, while S in (12) is hence a strictly concave function of R. We shall use these results in the following.

3.2 Censored residuals

When there are censored observations, a frequently used approach is to add the expected residual value to the censored residuals and then proceed as if one has a complete set of non-censored observations. For Cox-Snell residuals, the memory-less property of the exponential distribution implies that one should then add 1 to the censored residuals ((see e.g. Collett, 2003)). We will call these residuals the 1-adjusted Cox-Snell residuals.

It is interesting to note the connection between the 1-adjusted Cox-Snell residuals and what is known as martingale residuals, see e.g. Therneau et al. (2000) and Collett (2003, Ch. 4). Martingale residuals are given as

$$\hat{m}_i = \delta_i - \hat{r}_i.$$

Thus, since

$$\begin{aligned} 1 - \hat{m}_i &= \hat{r}_i \qquad \text{for non-censored observations} \\ 1 - \hat{m}_i &= \hat{r}_i + 1 \text{ for censored observations,} \end{aligned}$$

it is seen that martingale residuals, modulo a linear transformation, correspond to adding 1 to each $\hat{r}_i$ for a censored observation, which is exactly what the 1-adjusted Cox-Snell residuals do.

There is in the literature also an alternative adjusted Cox-Snell residual, which adds the amount $\log 2$ to the censored Cox-Snell residuals, corresponding to the median residual life of a unit exponentially distributed random variable. We will call them $\log 2$-adjusted Cox-Snell residuals. We shall see below that there are certain advantages with this convention when we deal with standardized residuals and Cox-Snell residuals in the same applications, as we will do in Section 5.

Consider an AFT model with a given distribution Φ for W. Then for a censored standardized residual s, all we know is that the "theoretical" standardized S as defined in (7) exceeds s. The 1-adjusted and $\log 2$-adjusted Cox-Snell residuals defined above, will for standardized residuals correspond to, respectively, replacing S by the expected value and the median of the conditional distribution of S given $S > s$, where S has distribution Φ.

Let now R be the "theoretical" Cox-Snell residual computed from S by (11). Then if $S > s$ we have

$$R > -\log(1 - \Phi(s)) \equiv r$$

The 1-adjusted Cox-Snell residual will now correspond to replacing r by $1 + r$, since

$$E(R|R > r) = 1 + r$$

But from this we have

$$1 + r = E[-\log(1 - \Phi(S)) \mid -\log(1 - \Phi(S)) > r] = E[-\log(1 - \Phi(S))|S > s],$$

which by the strict convexity of R as a function of S implies by Jensen's inequality that

$$1 + r > -\log(1 - \Phi(E(S|S > s)))$$

and hence that

$$\Phi^{-1}(1 - e^{1+r}) > E(S|S > s).$$

Now the left hand side of this inequality is the standardized residual corresponding to the 1-adjusted Cox-Snell residual, while the right hand side is the standardized residual $E(S|S > s)$. It is hence seen that 1-adjusted Cox-Snell residuals do not correspond to similar adjustments in the standardized residuals and vice versa.

This will however not be the case for the connection between the $\log 2$-adjusted Cox-Snell residual and the corresponding standardized residual based on the median. Let s be a censored standardized residual. Now replace it by the median of the conditional distribution of S given $S > s$, i.e. replace s by s' where

$$P(S > s'|S > s) = \frac{1}{2}.$$

By the strict monotonicity of R as a function of S this is equivalent to

$$P(R > -\log[1 - \Phi(s')] \mid R > -\log[1 - \Phi(s)]) = \frac{1}{2}.$$

Setting $r = -\log(1 - \Phi(s))$ it is clear from this that we can start by either of the censored residuals r and s and obtain the corresponding adjusted residual.

Most of our residual plotting methods, to be presented in the next sections, are based on exponential regression techniques, using both the non-adjusted and the adjusted Cox-Snell residuals, as well as the standardized adjusted residuals. For cases with a high degree of censoring it turns out that methods based on adjusted residuals may easily break down, so that the non-adjusted residuals should be preferred in this case. This will be the recommendation from several simulations. As mentioned in the introduction, we shall also consider residual plotting for discrete covariates.

4 Plots of residuals versus covariates

For each unit we observe a covariate vector $\boldsymbol{X}$. Let X be a specific component of this vector, and suppose that we will plot residuals versus this covariate. This corresponds of course to the standard procedure for residual plotting in ordinary linear regression. However, for plotting of residuals for censored survival data, it is clear that a plot of residuals versus covariate values may be misleading due the censored residuals being too small. It is because of this that the adjusted residuals have been introduced, and these may work well when there are not too many censored values. Crowley and Storer (1983) suggested, furthermore, in order to improve the symmetry of the residuals, to plot the logarithm of the Cox-Snell residuals. The logarithm of Cox-Snell residuals are then supposed to fluctuate around 0.

4.1 Continuous covariates

We shall adopt the idea of plotting the logarithm of Cox-Snell residuals, but for continuous covariates we shall impose some additional modeling and perform an exponential regression smoothing. This way of smoothing residuals will later turn out to be useful for the estimation of underlying covariate functions.

Let the data and the Cox-Snell residuals $\hat{r}_i$ be given as in the previous section. The idea is to consider a synthetic data set given as $(\hat{r}_1, \delta_1, x_1), \ldots, (\hat{r}_n, \delta_n, x_n)$, where $x_1, \ldots, x_n$ are the values of the specific covariate X for the n observation units, respectively, where we impose the following model for these data: $\hat{r}$ given x is exponentially distributed with hazard rate $\lambda(x)$, thus possibly depending on the covariate value. Then, based on the synthetic dataset, we use exponential regression to estimate the function $\lambda(\cdot)$, with estimate denoted $\hat{\lambda}(\cdot)$. In principle, any method for exponential regression can be used.

A residual plot versus x is then a plot of the estimated function $\log \hat{\lambda}(x)$, which may be revealed by the points $(x_i, \log \hat{\lambda}(x_i))$ for $i = 1, \ldots, n$. The idea is of course that if the assumed model is correct, then $\lambda(x)$ equals 1 for all x, so the $\hat{\lambda}(x_i)$ should be close to 1 and hence $\log(\hat{\lambda}(x_i))$ should fluctuate around 0.

There are several ways of performing the exponential regression, and we shall distinguish between two main classes of such methods. The first is for complete non-censored data, or for censored data with adjusted residuals for censored observations. For this class we suggest using local smoothers such as the loess (Cleveland, 1981) or other methods in the literature (see e.g. Hastie and Tibshirani, 1990).

The second class of methods apply to non-adjusted censored residuals. The so-called covariate order method (Kvaløy and Lindqvist, 2003, 2004) is tailored for this situation, as are certain Poisson regression methods (see e.g. Therneau et al., 2000). These methods have the advantage of working well for heavy censoring, where the methods mentioned above for adjusted residuals may break down. We will in particular use the covariate order method, for which there is also connected ways of testing of the null hypothesis of constant $\lambda(x)$ as suggested Kvaløy (2002) (see also Kvaløy and Lindqvist, 2003). A brief introduction to the covariate order method is given in Appendix C.

Example: Residual plots for Nelson's superalloy data

We consider an example from the book by Meeker and Escobar (1998), concerning the superalloy data from Nelson (1990).

The data give survival times measured in number of cycles, for 26 units of a superalloy, subject to different levels of pseudostress in a straincontrolled test. There is hence a single covariate in the model, and following Meeker and Escobar (1998) we shall consider the covariate to be $x = \log(\text{pseudostress})$.

The following model is fitted in Meeker and Escobar (1998),

$$\log Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \sigma W \tag{13}$$

where W is Gumbel distributed, i.e. Y is Weibull distributed.

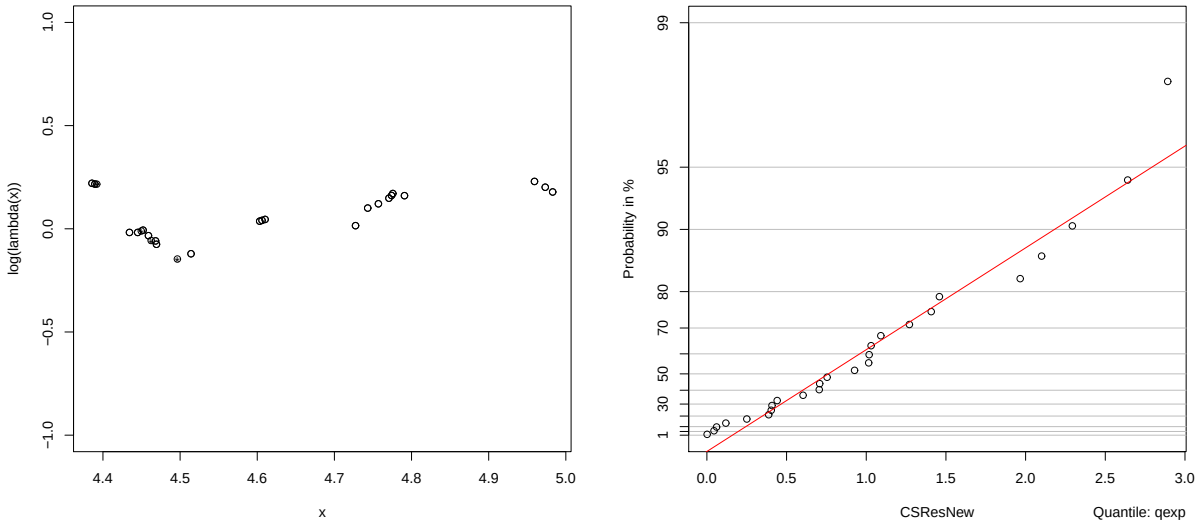

Figure 1: Superalloy data fitted with model (13). Left: Plot of logarithm of smoothed Cox-Snell residual $\log \hat{\lambda}(x)$ versus x. Circles represent failures, while dots correspond to censored events. Right: Exponential probability plot of 1-adjusted Cox-Snell residuals.

The left panel of Figure 1 shows the residual plot $(x_i, \log \hat{\lambda}(x_i))$, with $\hat{\lambda}(x)$ computed by the covariate order method (Kvaløy and Lindqvist, 2003). The plot indicates that the fluctuations from 0 are rather minor, and it is furthermore demonstrated by Aaserud (2011) that the discrepancy from a constant $\lambda(x)$ is not significant. On the other hand, the probability plot of the 1-adjusted Cox-Snell residuals shown in the right panel of Figure 1, which is supposed to be close to a straight line if the model is correct, indicates a slightly convex shape which may be due to a deficiency of the model. It is in this connection interesting to note that Meeker and Escobar (1998) suggest an extended AFT model where σ is allowed to depend on x.

4.2 Discrete covariates

Consider again the synthetic dataset $(\hat{r}_i, \delta_i, x_i)$; $i = 1, \ldots, n$, for a specific covariate X. Assuming that there are finitely many possible values for X, say k, we may divide the synthetic data into k sets and check each set for possible departures from a unit exponential distribution. This may be done for example by computing the estimated hazard rate $\lambda(x)$ for each possible value of x, under the assumption that residuals under x are exponential with hazard $\lambda(x)$. We may then (see example below) make separate probability plots for each possible value of X.

If the number of possible values for X is large, we may choose to smooth the estimated $\lambda(x)$ in a way similar to the continuous case. Alternatively, we may group values of X into a reasonable number of strata (see e.g. Arjas, 1988).

Example (Insulation data from Minitab)

The statistical package Minitab 16 (Minitab, Inc.) includes an example using data for deterioration of an insulation used for electric motors. This concerns an accelerated life time experiment where one wants to predict failure times for the insulation based on the temperature at which the motor runs.

The data give failure times for the insulation at four temperatures, 110, 130, 150, and 170 (degrees Celsius). The experiment is designed with 20 observations for each temperature, and there are altogether 14 right censored observations. Because the motors generally run at temperatures between 80 and 100 degrees, one wants to predict the insulation's behavior at those temperatures. It is thus important to have a good parametric model for the relationship between the temperature and the failure times.

We thus have a single covariate, temperature x, with four possible values, 110, 130, 150, 170. The following model is fitted in the Minitab application,

$$\log Y = \beta_0 + \gamma x + \sigma W, \tag{14}$$

where W has the Gumbel distribution. The fitted model is

$$\log Y = 16.2193 - 0.0572729x + 2.98957W.$$

from which we computed $\log 2$-adjusted Cox-Snell residuals. We computed Cox-Snell residuals, adding $\log 2$ to the censored ones. (We also tried the addition of 1, but the difference in final results was minor).

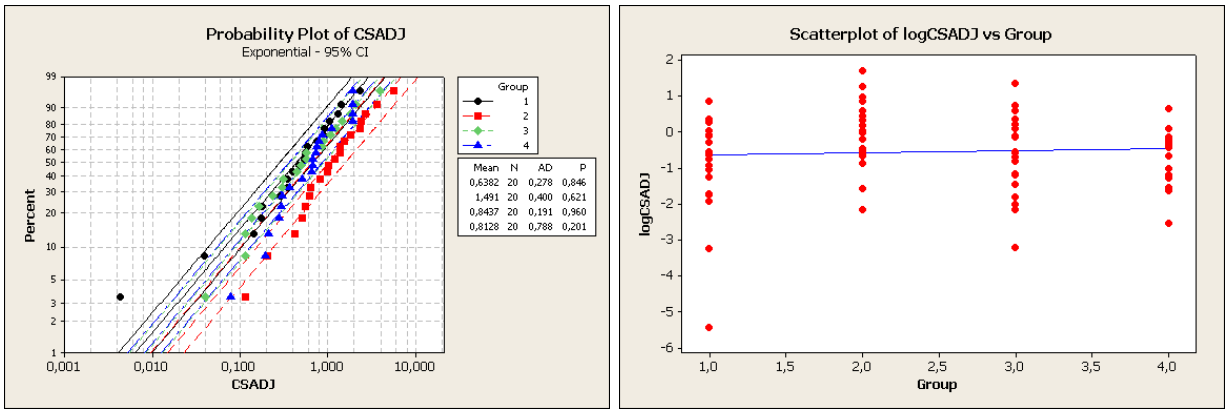

Figure 2: Insulation data fitted with (14). Left: Probability plots of adjusted Cox-Snell residuals for the four temperature groups. Right: The log of the $\log 2$-adjusted Cox-Snell residuals plotted against temperature.

The left panel of Figure 2 shows probability plots (with respect to exponential distribution) for each of the four temperatures (Group 1 = 170, Group 2 = 150 etc.), while the right panel shows log of the $\log 2$-adjusted Cox-Snell residuals plotted against group number. From the right diagram it seems that the distribution of the residuals from group 2 deviate from the

distributions of the residuals from groups 1,3,4, which appear to be closer to each other. The same effect is seen from the left diagram, where the points corresponding to group 2 form a curve close to a line, but clearly separated from the other groups. In fact, a Kruskal-Wallis test performed to compare the four groups of adjusted residuals, resulted in a p-value of 0.045, indicating a difference in the four distributions of residuals.

5 Functional form for a covariate

Suppose we want to conclude whether a specific covariate X, a component of the covariate vector $\boldsymbol{X}$ of the data, is appropriately represented in our model. This question may for example be triggered by a bad looking residual plot, obtained by one of the methods of the previous section. Alternatively, suppose that in the modeling process one starts by fitting a model with no covariates and then for each covariate X tries to find an appropriate covariate function $f(X)$ when X appears alone in the model. As yet another approach, one may, in an iterative manner, update one covariate function at a time, and derive covariate functions for all the covariates simultaneously, aiming at a representation (3).

In the present section we shall see how the above can be done by using the residuals and residual plots based on both standardized and Cox-Snell residuals.

5.1 Estimation of covariate functions

The following setup will essentially serve all the above mentioned possible procedures for derivation of covariate functions.

Assume that the correct model for the lifetime Y is

$$\log Y = \beta_0 + \boldsymbol{\beta}'\boldsymbol{Z} + f(X) + \sigma W, \tag{15}$$

where X is a single component of the vector $\boldsymbol{X}$, while $\boldsymbol{Z}$ is the vector of the remaining components of $\boldsymbol{X}$, so that $\boldsymbol{X} = (X, \boldsymbol{Z})$. Based on data $\{(t_i, \delta_i, x_i, \boldsymbol{z}_i); i = 1, \ldots, n\}$ we want to derive the appropriate form $f(X)$ for the covariate X.

Suppose we fit the simpler linear model

$$\log Y = \beta_0 + \boldsymbol{\beta}'\boldsymbol{Z} + \gamma X + \sigma W \tag{16}$$

by maximum likelihood estimation, using the likelihood function (6) with parameters defined by (16). Let the estimated model be

$$\log Y = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\boldsymbol{Z} + \hat{\gamma} X + \hat{\sigma} W.$$

Computation of standardized residuals using formula (8) then gives

$$\hat{s}_i = \frac{\log t_i - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}'\boldsymbol{z}_i - \hat{\gamma} x_i}{\hat{\sigma}}.$$

Recall that these will approximately behave like observations with distribution Φ if (16) is the *correct* model, that is if $f(x)$ in fact is linear in x; while if this is not the case, the

residuals may behave quite differently. It is the purpose of the following to show how the $\hat{s}_i$ can be used to infer the true form of $f(x)$.

The clue is a result by White (1982) on maximum likelihood estimation in misspecified parametric models. It follows from White (1982) that, under appropriate conditions, there are parameter values $(\beta_0^*, \boldsymbol{\beta^*}, \gamma^*, \sigma^*)$ of the possibly wrong model (16) which are the limits (a.s.) of the estimators $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\sigma})$ as $n \to \infty$. The $(\beta_0^*, \boldsymbol{\beta^*}, \gamma^*, \sigma^*)$ are, more precisely, given as the minimizers of the Kullback-Leibler distance between the true model as defined by (15) and the possibly misspecified model (16).

Appendix B derives in some special cases the expressions that are minimized in order to find the starred parameters, and solves the minimization problem analytically in certain simple cases. An example of how to find the starred parameters by simulation is also given.

In the model defined by $(\beta_0^*, \boldsymbol{\beta^*}, \gamma^*, \sigma^*)$ we would compute the "theoretical" standardized residual from (9) as

$$S^* = \frac{\log Y - \beta_0^* - \boldsymbol{\beta^{*\prime}} \boldsymbol{Z} - \gamma^* X}{\sigma^*},$$

which by inserting the true model for $\log Y$, (15), can be written

$$S^* = \frac{\sigma}{\sigma^*} W + \frac{(\beta - \beta_0^*) + (\boldsymbol{\beta} - \boldsymbol{\beta^*})' \boldsymbol{Z} + f(X) - \gamma^* X}{\sigma^*}. \tag{17}$$

It should be clear that if $f(X)$ really is linear, then S^*, conditional on $\boldsymbol{X} = (X, \boldsymbol{Z})$ is exactly distributed as W.

Solving (17) for $f(X)$ gives

$$f(X) = -\sigma W - (\beta - \beta_0^*) - (\boldsymbol{\beta} - \boldsymbol{\beta^*})' \boldsymbol{Z} + \gamma^* X + \sigma^* S^*, \tag{18}$$

Now, take the conditional expectation given $X = x$, throughout the equation (18), and recall that X is the first component of $\boldsymbol{X} = (X, \boldsymbol{Z})$. This gives

$$f(x) = -\sigma E(W) - (\beta - \beta_0^*) - (\boldsymbol{\beta} - \boldsymbol{\beta^*})' E(\boldsymbol{Z} | X = x) + \gamma^* x + \sigma^* E(S^* | X = x). \tag{19}$$

The rest of the present section is concerned with the practical use of the result (19). Assume first that X and $\boldsymbol{Z}$ are independent. Then (19) implies that $f(x)$ is of the form

$$f(x) = \text{constant} + \gamma^* x + \sigma^* E(S^* | X = x), \tag{20}$$

where the constant does not depend on x. Thus, based on our data we obtain that, modulo an unknown additive constant, we can estimate the function $f(x)$ by

$$\hat{f}(x) = \hat{\gamma} x + \hat{\sigma} \hat{H}(x), \tag{21}$$

where $\hat{H}(x)$ is an estimate of

$$H(x) \equiv E(S^* | X = x). \tag{22}$$

If there are no censorings, or if we use adjusted standardized residuals $\hat{s}_i$, then the function $H(x)$ can be estimated by smoothing the points $(x_i, \hat{s}_i)$; $i = 1, \ldots, n$. For estimation of $H(x)$ from non-adjusted residuals, we refer to Section 5.3.

It follows from (19) that if $\boldsymbol{Z}$ and X are dependent, then the $\hat{f}(x)$ in (21) may be somewhat "blurred", with a degree of blurring depending on the kind and amount of dependency between $\boldsymbol{Z}$ and X. On the other hand, if (16) is approximately true, then $\boldsymbol{\beta} - \boldsymbol{\beta^*}$ may be expected to be close to the zero vector, so the dependency of $\boldsymbol{Z}$ and X may not influence $\hat{f}(x)$ seriously. Examples of analytical computation of $\boldsymbol{\beta} - \boldsymbol{\beta^*}$ for particular cases is given in Appendix B.

5.2 Functional form for covariates for non-censored or adjusted residuals

In this subsection we consider estimation of the function $H(x)$ when either all residuals are non-censored, or the censored residuals are adjusted as described earlier. Suppose first that X is a discrete covariate with a finite number of possible values x. Let $p(x) = P(X = x)$, and let us act as if there are no censored observations.

A natural estimator for $H(x)$, when $p(x) > 0$, from observations $(Y_i, X_i, \boldsymbol{Z}_i)$, is

$$\hat{H}(x) = \frac{\sum_{i:X_i=x} \hat{S}_i}{n(x)}, \tag{23}$$

where $n(x) = \#\{i : X_i = x\}$ and

$$\hat{S}_i = \frac{\log Y_i - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}' \boldsymbol{Z_i} - \hat{\gamma} X_i}{\hat{\sigma}}.$$

Now there are underlying variables $W_1, \ldots, W_n$ such that $\log Y_i = \beta_0 + \boldsymbol{\beta}' \boldsymbol{Z_i} + f(X_i) + \sigma W_i$ for $i = 1, \ldots, n$, and hence we can write

$$\hat{S}_i = \frac{(\beta_0 - \hat{\beta}_0) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \boldsymbol{Z_i} + f(X_i) - \hat{\gamma} X_i + \sigma W_i}{\hat{\sigma}}.$$

From this,

$$\hat{H}(x) = \frac{(\beta_0 - \hat{\beta}_0) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' (\sum_{i:X_i=x} \boldsymbol{Z_i}/n(x)) + f(x) - \hat{\gamma} x + \sigma \sum_{i:X_i=x} W_i/n(x)}{\hat{\sigma}}.$$

Since the W_i are independent of the X_i, it is clear by SLLN that $\sum_{i:X_i=x} W_i/n(x)$ converges (a.s.) to $E(W)$. Further, the $\boldsymbol{Z}_i$ for $i \in \{i : X_i = x\}$ is clearly a sample from the conditional distribution of $\boldsymbol{Z}$ given $X = x$, and hence by SLLN converges to $E(\boldsymbol{Z}|X = x)$. Since $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\sigma}) \to (\beta_0^*, \boldsymbol{\beta}^*, \sigma^*)$ (a.s.) it follows that $\hat{H}(x) \to H(x)$ for all x with $p(x) > 0$. This proves that $\hat{H}(x)$ as defined in (23) is a strongly consistent estimator for $H(x)$ to be used in (21).

If X is a continuous covariate, we may extend the proof by considering partitions of the range of X and taking appropriate limits as n tends to infinity and partitions become finer and finer. This may be used to prove convergence of smoothing procedures based on the points $(x_i, \hat{s}_i)$ as suggested in the previous subsection. It is, however, beyond the scope of this paper to do this in detail.

It should be remarked that the use of adjusted residuals may be inefficient. This is because we modify censored residuals by assuming that they have the distribution Φ, while as we have seen in Section 5.1, it is the deviance of the residuals from their standard distribution that makes the method work! In the case of many censorings, we may hence get misleading results. In the case of medium to high censoring, we will advocate methods which treat the censored residuals more appropriately. We return to this in the next subsection (Section 5.3), and a simulation example will be given in Section 6.

Example (Insulation data from Minitab, continued)

Based on the insulation data we would like to reconsider the functional form for the covariate X. We then compute $\hat{H}(x)$ for each of the four groups, using formula (23), obtaining respective values -.9902,.0293,-.7315, -.5263.

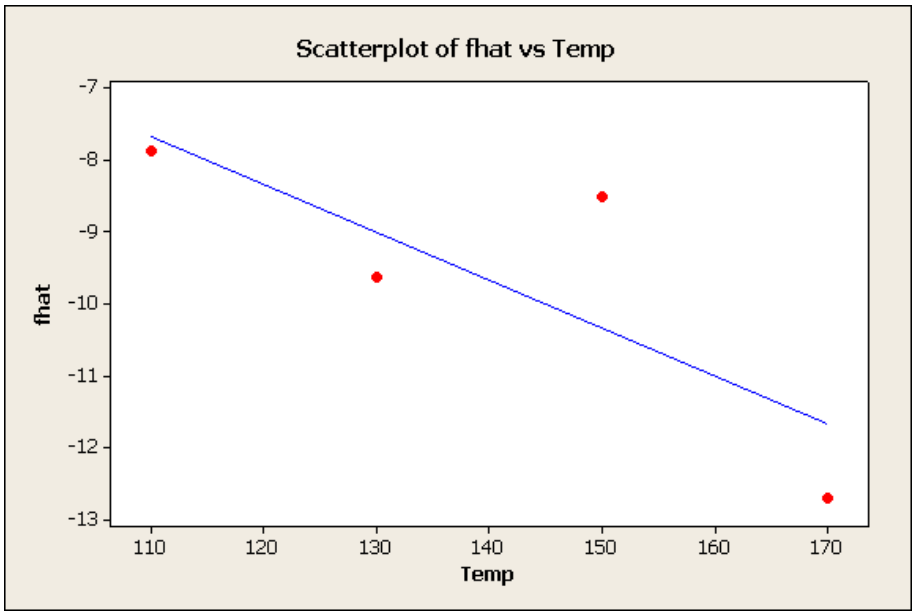

Figure 3: The four estimates $\hat{f}(x)$ plotted against temperature. The line is the least squares line based on the four points.

Using the formula (21) we then get

$$\begin{aligned}
\hat{f}(170) &= -.0572729 \cdot 170 - 2.98957 \cdot .9902 = -12.69659 \\
\hat{f}(150) &= -.0572729 \cdot 150 + 2.98957 \cdot .0293 = -8.503447 \\
\hat{f}(130) &= -.0572729 \cdot 130 - 2.98957 \cdot .7315 = -9.632198 \\
\hat{f}(110) &= -.0572729 \cdot 110 - 2.98957 \cdot .5263 = -7.873295
\end{aligned}$$

The $\hat{f}$ is graphed in Figure 3, together with the least squares line. The model that was suggested in the first part of this example (Section 4.2) corresponds to a covariate function

given by a straight line in this diagram. The resulting $\hat{f}$ clearly deviates from a line, however, a fact which is consistent with the result of the residual analysis in Section 4.2. As seen there, the results from group 2 (150 degrees) cause an apparent deviation from the originally assumed model.

Noting that the aim of the experiment behind the data, was to extrapolate properties of the insulation material to temperatures between 80 and 100 degrees, the conclusions obtained here should presumably lead to further investigation and experimentation.

5.3 Functional form for covariates based on smoothing of residuals

We have already seen (Section 4) how the non-adjusted residuals can be used via censored exponential regression to estimate nonparametrically the hazard $\lambda(x)$ of the Cox-Snell residual corresponding to covariate x. The estimated functions $\hat{\lambda}(x)$ can now be used in (21) if we replace $\hat{H}(x)$ by the estimate

$$\hat{\hat{H}}(x) = \Phi^{-1}(1 - \exp(-1/\hat{\lambda}(x))). \tag{24}$$

The reason for this is that by (12), (22) can be written in terms of the Cox-Snell residual R^* as

$$H(x) \equiv E(\Phi^{-1}(1 - \exp(-R^*))|X = x). \tag{25}$$

Hence (24) is established by replacing R^* in (25) for a given x by its estimated expected value.

It follows from Appendix A that the function $r \mapsto \Phi^{-1}(1 - e^{-r})$ is concave for the commonly considered models, lognormal, Weibull and log-logistic. But then the right hand side of (24) is convex in $\hat{\lambda}(x)$, so by Jensen's inequality, $E(\hat{\hat{H}}(x)) \geq \Phi^{-1}(1-\exp(-1/E(\hat{\lambda}(x)))$ which indicates a possibility of overestimation. The practical consequences of this convexity is, however, not clear. In any case, if $\hat{\lambda}(x)$ consistently estimates $\lambda(x)$, then $\hat{\hat{H}}(x)$ estimates $H(x)$ consistently under the given assumptions.

In the next section we consider in more detail the Weibull case, which is probably the most used model in practice.

6 Weibull AFT models

Suppose now that T is Weibull-distributed, and hence that W has the Gumbel distribution, with

$$\Phi(u) = 1 - e^{-e^u} \text{ for } -\infty < u < \infty$$

The Cox-Snell residuals (10) and the standardized residuals (8) are hence given from

$$\hat{r}_i = \exp\left(\frac{\log t_i - \hat{f}(\boldsymbol{x}_i)}{\hat{\sigma}}\right) = e^{\hat{s}_i}.$$

From (17) follows that the theoretical Cox-Snell residual, computed from the misspecified model (16), can be written as

$$R^* = U^{\sigma/\sigma^*} \exp\left(\frac{(\beta - \beta_0^*) + (\boldsymbol{\beta} - \boldsymbol{\beta^*})'\boldsymbol{Z} + f(X) - \gamma^* X}{\sigma^*}\right), \tag{26}$$

where $U = e^W$ is unit exponentially distributed. This shows that R^* under the true model is in fact conditionally Weibull-distributed with shape parameter σ^*/σ and scale parameter

$$\exp\left(\frac{(\beta - \beta_0^*) + (\boldsymbol{\beta} - \boldsymbol{\beta^*})'\boldsymbol{Z} + f(X) - \gamma^* X}{\sigma^*}\right).$$

This is an interesting observation, since the exponential regression methods we have suggested in Section 5.3 assume that R^* is (approximately) exponentially distributed. The practical problem is of course that while σ^* is consistently estimable by $\hat{\sigma}$ we can not estimate σ since we do not know the true model. Thus we are not able to estimate σ/σ^* and essentially we then decide to set it to 1 in our approach.

Note further that for the Weibull case,

$$\Phi^{-1}(x) = \log(-\log(1-x)) \text{ for } 0 < x < 1,$$

so $H(x) = E(\log R^* | X = x)$, which when using adjusted residuals can be estimated by smoothing the points $(x_i, \log \hat{r}_i)$. Recalling that the martingale residuals are given by $\hat{m}_i = 1 - \hat{r}_i$ it follows from this that $H(x)$ can be estimated by smoothing the points $(x_i, \log(1-\hat{m}_i))$ or approximately the points $(x_i, -\hat{m}_i))$. Thus for the case where we start by fitting a model without the γx term (see (16)), we obtain the shape of $-f(x)$ by plotting the martingale residuals versus the x-values. This corresponds to the approach suggested in Therneau et al. (1990) for Cox-regression.

Furthermore, it follows that $\hat{\hat{H}}(x)$ in (24) has the simple form

$$\hat{\hat{H}}(x) = -\log \hat{\lambda}(x). \tag{27}$$

which by (21) gives the useful formula

$$\hat{f}(x) = \hat{\gamma} x - \hat{\sigma} \log \hat{\lambda}(x) \tag{28}$$

(modulo a constant).

Example (Alloy data, continued)

Suppose that we start by fitting the empty Weibull model, i.e. the model

$$\log Y = \beta_0 + \sigma W \tag{29}$$

where W is Gumbel distributed. It follows that we may use

$$\hat{f}(x) = \hat{\beta}_0 - \log \hat{\lambda}(x), \tag{30}$$

where $\hat{\lambda}(x)$ is the covariate order smoothing of the Cox-Snell residuals resulting from fitting model (29).

The resulting curve is shown in Figure 4. Examining the plot it seems rather clear that a linear function for $f(x)$ is not reasonable, and as we have already seen in this example, a quadratic function gives a satisfactory fit. This might be suggested by Fig 4, and this demonstrates one possible application of our approach.

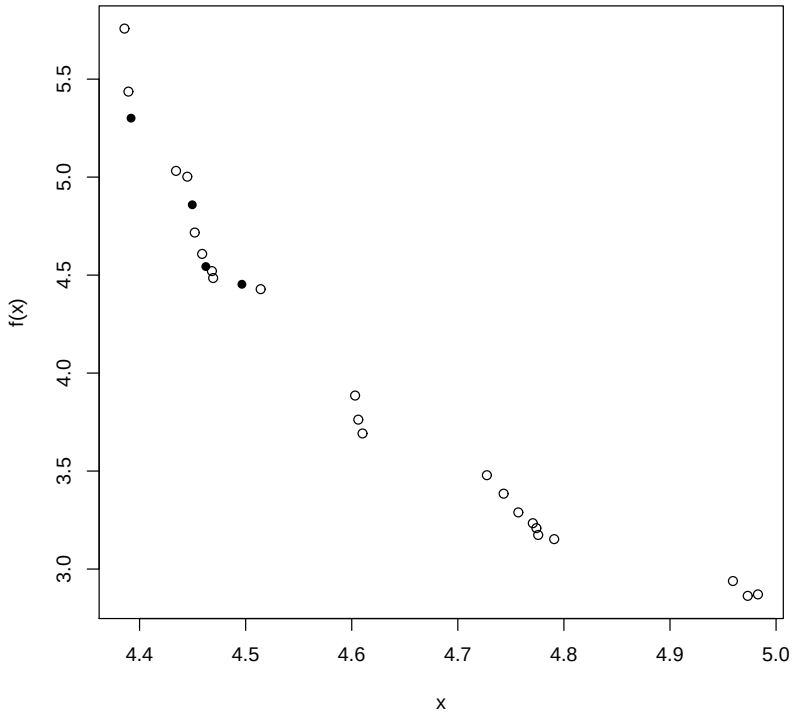

Figure 4: Superalloy data fitted with empty model, i.e. $\log Y = \beta_0$. The plot shows $\hat{f}(x)$ obtained by (30) using the covariate order method

Example (Simulated data from Weibull-distribution):

We simulated $n = 100$ observations from the Weibull-distribution using the model

$$\log Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + f(X) + \sigma W,$$

where $\beta_0 = 0,\ \beta_1 = 5,\ \beta_2 = 0.2,\ f(x) = x^2,\ \sigma = 2$; the W were drawn from the Gumbel distribution, while the Z_1, Z_2, X were independently drawn from standard normal distributions. We imposed two different censoring scenarios by drawing independent censoring times C giving approximately 20% and 50% censoring, respectively.

Figure 5 shows the resulting estimates of the true covariate function $f(X) = X^2$, using both a loess smoothing on the adjusted residuals, and a censored nonparametric exponential regression using the nonadjusted residuals. A possible conclusion from this and similar datasets is that there are no large differences between the two methods for estimation of the covariate function $f(X)$ for low censoring, while for more heavy censoring the nonparametric exponential regression method seemingly performs slightly better.

7 Proportional hazards models

Therneau et al. (1990) and Grambsch et al. (1995) considered the derivation of covariate func-

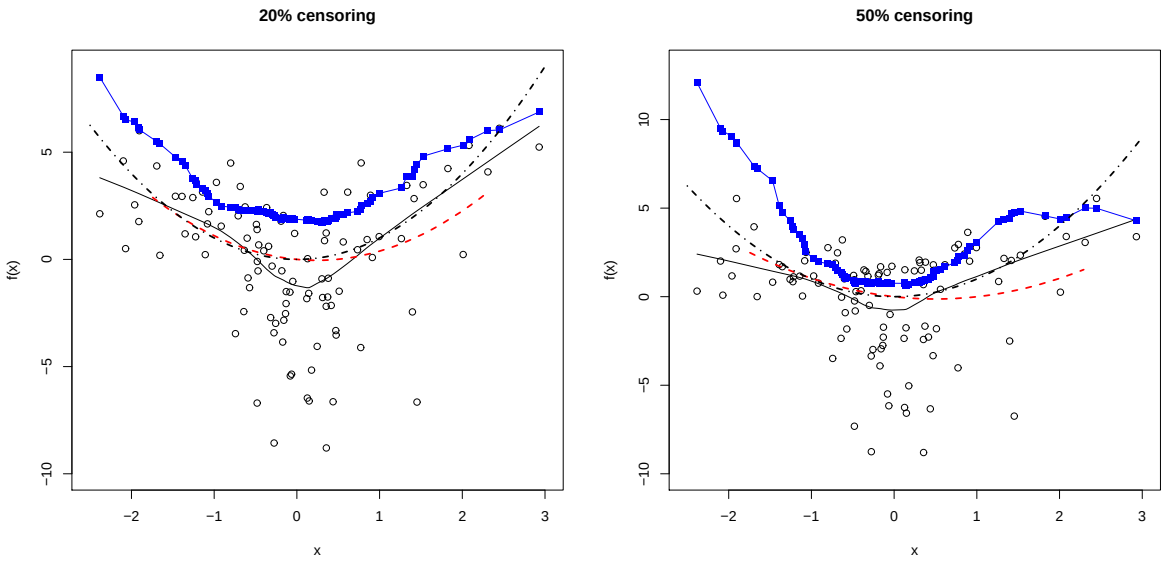

Figure 5: Simulated Weibull distributed data. Circles are $(x_i, \hat{\gamma} x_i + \hat{\sigma} \log \hat{r}_i)$ using the 1-adjusted Cox-Snell residuals; solid line is loess smooth from these points; dashed line is a quadratic function fitted to the same points; squares are the $\hat{f}(x_i)$ obtained from (28). The true quadratic curve is given by dash-dots.

tions in Cox' proportional hazards model. The former paper touched in the last paragraph the problem of parametric proportional hazards models, but without giving any explicit results. In the present section we shall briefly consider this type of models and see how the approach for AFT models can be modified for such cases.

As is well known (see e.g. Cox and Oakes, 1984), the only AFT models which are proportional hazards models are the Weibull models. We shall here more generally consider the parametric proportional hazards model where the lifetime Y conditional on the covariate vector $\boldsymbol{X}$ has hazard function

$$\lambda(t|\boldsymbol{X}) = g(t, \boldsymbol{\theta}) \exp\{\beta_0 + \boldsymbol{\beta}' \boldsymbol{X}\}. \tag{31}$$

Here $\boldsymbol{\theta}$ is a parameter vector for the baseline hazard function, while β_0 and $\boldsymbol{\beta}$ are unknown regression coefficients. The difference from a Cox model is hence the parametric form of the baseline hazard. Note that we include an intercept term β_0 in the linear function of the covariates. By this we avoid the need for a scale factor in the baseline hazard $g(t, \boldsymbol{\theta})$. For example, a Weibull regression model can be represented by $g(t, \theta) = \theta t^{\theta - 1}$ for $\theta > 0$, while a Gompertz model (see Collett, 2003, p. 191) may have $g(t, \theta) = e^{\theta t}$ for $-\infty < \theta < \infty$.

Now suppose we would like to infer the appropriate covariate function $f(X)$ for a single covariate X. We will use a similar approach as has been used for the AFT models, so only the crucial steps are included below. Let $\boldsymbol{X} = (X, \boldsymbol{Z})$ and assume that the true model has hazard rate

$$\lambda(t|\boldsymbol{X}) = g(t, \boldsymbol{\theta}) \exp\{\beta_0 + \boldsymbol{\beta}' \boldsymbol{Z} + f(X)\}.$$

By fitting the model

$$\lambda(t|\boldsymbol{X}) = g(t, \boldsymbol{\theta}) \exp\{\beta_0 + \boldsymbol{\beta}' \boldsymbol{Z} + \gamma X\}$$

by maximum likelihood estimation our estimator $(\hat{\boldsymbol{\theta}}, \hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\gamma})$ will be a consistent estimator for $(\boldsymbol{\theta}^*, \hat{\beta}_0^*, \boldsymbol{\beta}^*, \gamma^*)$, say.

The "theoretical" Cox-Snell residuals based on the fitted model are

$$R^* = G(Y, \boldsymbol{\theta}^*) \exp\{\beta_0^* + \boldsymbol{\beta}^{*'} \boldsymbol{Z} + \gamma^* X\},$$

where $G(t, \boldsymbol{\theta}) = \int_0^t g(u, \boldsymbol{\theta}) du$ is the integrated baseline hazard function.

Now we compute

$$\begin{aligned}
E(\log R^* | X) &= E[\log G(Y, \boldsymbol{\theta}^*) + \beta_0^* + \boldsymbol{\beta}^{*'} \boldsymbol{Z} + \gamma^* X | X] \\
&= E[\log G(Y, \boldsymbol{\theta}^*) - \log G(Y, \boldsymbol{\theta}) \\
&\quad + \log G(Y, \boldsymbol{\theta}) + \beta_0 + \boldsymbol{\beta}' \boldsymbol{Z} + f(X) \\
&\quad - \beta_0 - \boldsymbol{\beta}' \boldsymbol{Z} - f(X) + \beta_0^* + \boldsymbol{\beta}^{*'} \boldsymbol{Z} + \gamma^* X | X] \\
&= E[\log \frac{G(Y, \boldsymbol{\theta}^*)}{G(Y, \boldsymbol{\theta})} | X] \\
&\quad + E\{E[\log G(Y, \boldsymbol{\theta}) + \beta_0 + \boldsymbol{\beta}' \boldsymbol{Z} + f(X) | X, \boldsymbol{Z}]\} \\
&\quad + \beta_0^* - \beta_0 + (\boldsymbol{\beta}^* - \boldsymbol{\beta})' E[\boldsymbol{Z} | X] + \gamma^* X - f(X) \\
&= E[\log \frac{G(Y, \boldsymbol{\theta}^*)}{G(Y, \boldsymbol{\theta})} | X] - a + \beta_0^* - \beta_0 + (\boldsymbol{\beta}^* - \boldsymbol{\beta})' E[\boldsymbol{Z} | X] + \gamma^* X - f(X),
\end{aligned}$$

where $a = 0.577215665\ldots$ is Euler's constant. Here we have used that

$$R = G(Y, \boldsymbol{\theta}) \exp\{\beta_0 + \boldsymbol{\beta}' \boldsymbol{Z} + f(X)\}$$

is unit exponentially distributed, conditionally on $X, \boldsymbol{Z}$, so that $\log R$ is standard Gumbel distributed, again conditionally on $X, \boldsymbol{Z}$.

Under the assumption that either $\boldsymbol{Z}$ is independent of X, or $\boldsymbol{\beta}^* \approx \boldsymbol{\beta}$, and furthermore assuming that $E[\log \frac{G(Y, \boldsymbol{\theta}^*)}{G(Y, \boldsymbol{\theta})} | X]$ varies slowly with X, the above computation indicates how the log of Cox-Snell residuals can be used to infer the form of appropriate covariate functions $f(X)$. The practical use of the result is now much similar to the approach considered in Section 5 and is not considered further here.

The approximations suggested above correspond to the approximations used by Therneau et al. (1990) for Cox regression, and these were also used by Kvaløy and Lindqvist (2003). Note that the approximations essentially amount to assuming that the cumulative baseline hazard does not change much under models that are close together.

Example: Weibull regression

Now suppose $g(t, \theta) = \theta t^{\theta - 1}$, so $G(t, \theta) = t^\theta$. Then it can be shown that

$$E[\log \frac{G(Y, \boldsymbol{\theta}^*)}{G(Y, \boldsymbol{\theta})} | X] = (1 - \frac{\theta^*}{\theta})(\beta_0 + \boldsymbol{\beta}' E(\boldsymbol{Z} | X) + f(X) + a)$$

so that

$$E(\log R^* | X) = \beta_0^* - \frac{\theta^*}{\theta} \beta_0 + (\boldsymbol{\beta}^* - \frac{\theta^*}{\theta} \boldsymbol{\beta})' E[\boldsymbol{Z} | X] + \gamma^* X - \frac{\theta^*}{\theta} f(X) - \frac{\theta^*}{\theta} a$$

If $\boldsymbol{Z}$ and X are independent, then this equals a constant plus the term $\gamma^* X - \frac{\theta^*}{\theta} f(X)$. Similarly to what we have seen for the AFT case, here θ^* and γ^* can be consistently estimated by $\hat{\theta}$ and $\hat{\gamma}$, while θ can not be estimated. Thus for practical purposes we assume that $\theta^*/\theta = 1$, which is believed to work well in order to recover the basic shape of $f(X)$. The result here is consistent with that of Section 6, where the apparent difference caused by $\gamma^* X - \frac{\theta^*}{\theta} f(X)$ appearing in the present case instead of simply $\gamma^* X - f(X)$, is due to the difference in the way the linear function of the covariates is included in the AFT and proportional hazard models.

Example: Gompertz regression

For this model we have

$$G(t, \theta) = \frac{e^{\theta t} - 1}{\theta}$$

From the approximation (valid for small $|\theta|$) $G(t, \theta) \approx t + (1/2)\theta t^2$, we get

$$E[\log \frac{G(Y, \boldsymbol{\theta}^*)}{G(Y, \boldsymbol{\theta})} | X] \approx \frac{1}{2}(\theta^* - \theta) E(Y|X),$$

which will be small if θ^* is close to θ.

8 Discussion and conclusion

It has been shown how residuals from AFT models can be obtained and plotted, also in cases with a large amount of censored observations. For continuous covariates, various smoothing techniques have been suggested. In cases where the residual plots are not satisfactory, it is furthermore demonstrated how the computed residuals can be used to improve the model by suggesting appropriate functions of the residuals to be used in the AFT model. These techniques can also be used to build an AFT model step by step by introducing one covariate at a time.

We remark that in the parametric AFT model (1) the "error" is represented by a single parameter σ, in addition to the W having a known distribution. In the parametric proportional hazards model (31), on the other hand, we allowed a multidimensional parameter $\boldsymbol{\theta}$ of the baseline hazard function (although the two examples had just one parameter in the baseline). For the AFT approach, the analogue to the multiparameter baseline would be to include unknown parameters into the distribution of W. The corresponding modifications of the methods of the present paper appear rather straightforward in a maximum likelihood approach. Another extension of the considered AFT models would be to let also σ depend on the covariates. Such a possibility is in fact already mentioned in connection with the data example of Section 4.1, and was suggested in an example of Meeker and Escobar (1998).

References

Aaserud, S. (2011). Residuals and functional form in accelerated life regression models. Master's thesis, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim.

Andersen, P. K., Borgan, Ø., Gill, R. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *Journal of the American Statistical Association*, 83, 204–212.

Cleveland, W. S. (1981). Lowess: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, 35, 54.

Collett, D. (2003). *Modelling Survival Data in Medical Research.* Chapman & Hall, London.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman & Hall, London.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion). *Journal of Royal Statistical Society, Ser. B.*, 30, 248–75.

Crowley, J. and Storer, B. E. (1983). A reanalysis of the Stanford heart transplant data: Comment. *Journal of the American Statistical Association*, 78, 277–281.

Grambsch, P. M., Therneau, T. M. and Fleming, T. R. (1995). Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. *Biometrics*, 51, 1469–1482.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman & Hall, London.

Kay, R. (1977. Proportional hazard regression models and the analysis of censored survival data. *Applied Statistics*, 26, 227–237.

Kvaløy, J. T. (2002). Covariate order tests for covariate effect. *Lifetime Data Analysis*, 8, 35–52.

Kvaløy, J. T. and Lindqvist, B. H. (2003). Estimation and inference in nonparametric Cox-models: Time transformation methods. *Computational Statistics*, 18, 205–221.

Kvaløy, J. T. and Lindqvist, B. H. (2004). The covariate order method for nonparametric exponential regression and some applications in other lifetime models. In M. S. Nikulin, N. Balakrishnan, M. Mesbah and N. Limnios (eds). *Parametric and Semipara- metric Models with Applications to Reliability, Survival Analysis, and Quality of Life.* Birkhauser, pp. 221–237.

Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data.* Wiley, New York.

Nelson, W. (1990). *Accelerated Testing: Statistical Models, Test Plans, and Data Analyses.* Wiley, N.Y.

Silverman, B. W. (1986). *Density Estimation* Chapman & Hall, London.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model.* Springer-Verlag, New York.

Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77, 147–160.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Appendix

A On the relation between standardized residuals and Cox-Snell residuals

The following general result, easily proved by differentiation, can be used to determine convexity and concavity of the functions (11)-(12) for a given distribution Φ.

Lemma 1 *The function $s \mapsto -\log(1-\Phi(s))$ is strictly convex (and hence the inverse function $r \mapsto \Phi^{-1}(1-e^{-r})$ for $r > 0$ is strictly concave) if and only if*

$$\Phi''(u)(1-\Phi(u)) + (\Phi'(u))^2 > 0$$

for all $-\infty < u < \infty$.

It is easy to see that the condition holds for the Weibull and log-logistic cases, for which we have, respectively, $\Phi(u) = 1 - e^{-e^u}$ and $\Phi(u) = e^u/(1+e^u)$.

Now we shall see that the lemma holds also when Φ is the standard normal distribution. Let $\phi(u)$ be the density function of the standard normal distribution, so $\Phi'(u) = \phi(u)$ and hence $\Phi''(u) = -u\phi(u)$. Thus, the condition in the lemma is equivalent to

$$u(1-\Phi(u)) < \phi(u)$$

for all u. This is of course trivial for $u \leq 0$. For $u > 0$ it is equivalent to $1-\Phi(u)-\phi(u)/u < 0$. That this holds for all $u > 0$ is seen by showing that the expression is increasing in u and tends to 0 as $u \to \infty$, which is fairly straightforward.

B More on parameters of misspecified models

Following White (1982), the starred parameters defined in Section 5 are found by minimizing the expected value of the log of the ratio between the density for $(T, \Delta, \boldsymbol{X})$ under the true model and under the misspecified model, when the random variables themselves, having the true distribution, are used in the densities. The densities to be used are hence (5). It is here natural to assume that the distribution of $\boldsymbol{X}$ and the conditional distribution of C given $\boldsymbol{X}$ are the same for both models, so we compare in effect the densities

$$g_Y(t|\boldsymbol{x})^{\delta} G_Y(t|\boldsymbol{x})^{1-\delta} \tag{32}$$

corresponding to the two models.

Under the assumptions for AFT models, the general expressions for the functions in (32) are

$$\begin{aligned} g_Y(t|\boldsymbol{x}) &= \frac{1}{t\sigma}\phi\left(\frac{\log t - f(\boldsymbol{x})}{\sigma}\right) \\ G_Y(t|\boldsymbol{x}) &= 1 - \Phi\left(\frac{\log t - f(\boldsymbol{x})}{\sigma}\right) \end{aligned}$$

From this, and substituting the assumed form of $f(\boldsymbol{x})$ for the two models, we can write the criterion to be minimized as

$$\begin{aligned} E(D) &\equiv E\left[\Delta \log \frac{\frac{1}{\sigma}\phi\left(\frac{\log T-\beta_0-\boldsymbol{\beta}'Z-f(X)}{\sigma}\right)}{\frac{1}{\sigma^*}\phi\left(\frac{\log T-\beta_0^*-\boldsymbol{\beta}^{*\prime}Z-\gamma^* X}{\sigma^*}\right)} + (1-\Delta)\log\frac{1-\Phi\left(\frac{\log T-\beta_0-\boldsymbol{\beta}'Z-f(X)}{\sigma}\right)}{1-\Phi\left(\frac{\log T-\beta_0^*-\boldsymbol{\beta}^{*\prime}Z-\gamma^* X}{\sigma^*}\right)}\right] \\ &= E\left[\Delta \log\left\{\frac{\sigma^*}{\sigma}\frac{\phi(W)}{\phi\left(\frac{\sigma}{\sigma^*}W + \frac{(\beta_0-\beta_0^*)+(\boldsymbol{\beta}-\boldsymbol{\beta^*})'\boldsymbol{z}+f(x)-\gamma^* x}{\sigma^*}\right)}\right\}\right] \\ &+ E\left[(1-\Delta)\log\frac{1-\Phi(W)}{1-\Phi\left(\frac{\sigma}{\sigma^*}W + \frac{(\beta_0-\beta_0^*)+(\boldsymbol{\beta}-\boldsymbol{\beta^*})'\boldsymbol{z}+f(x)-\gamma^* x}{\sigma^*}\right)}\right] \end{aligned} \tag{33}$$

where expectation is taken with respect to the true joint distribution of $(\boldsymbol{X}, W, C)$, where $\boldsymbol{X} = (X, \boldsymbol{Z})$. Note also that we can express Δ in terms of $(\boldsymbol{X}, W, C)$ as $\Delta = I(\beta_0 + \boldsymbol{\beta}'\boldsymbol{Z} + f(X) + \sigma W < C)$. The task is to minimize the expression in (33) with respect to the starred parameters. From the expression it can be seen that the minimizing parameters may depend on the censoring distribution, which is an interesting observation since it is well known that for a correctly specified distribution, the maximum likelihood estimators under independent censoring will always converge to the true model independently of the censoring scheme.

In general the minimization of (33) may be difficult to do analytically. A simple way of "cheating" to get approximate values for the starred parameters is to simulate from the true model a (very) large number of observations and then use a statistical package (e.g. R) to compute the maximum likelihood estimators. This has been done by Aaserud (2011), see example below. First we shall for illustration go through some examples of how (33) will look in particular cases, and in some cases we also show how it can be minimized analytically.

Example - lognormal distribution

Assume here that W has the standard normal distribution, and that there is no censoring. Assume for simplicity that Z is one-dimensional and assume without loss of generality that $E(X) = E(Z) = 0$.

From (33) with $P(\Delta = 1) = 1$, with ϕ being the standard normal density and W being standard normally distributed, we have

$$E(D) = \log \frac{\sigma^*}{\sigma} - \frac{1}{2} + \frac{\sigma^2}{2\sigma^{*2}} + \frac{E\{(\beta_0 - \beta_0^* + (\beta_1 - \beta_1^*)Z + f(X) - \gamma^* X)^2\}}{2\sigma^{*2}}.$$

By differentiation with respect to all the starred parameters we obtain the solutions

$$\begin{aligned}
\beta_0^* - \beta_0 &= Ef(X) \\
\beta_1^* - \beta_1 &= \frac{E\{Zf(X)\} - E\{XZ\}E\{Xf(X)\}}{1 - (E\{XZ\})^2} \\
\gamma^* &= \frac{E\{Xf(X)\} - E\{XZ\}E\{Zf(X)\}}{1 - (E\{XZ\})^2} \\
\sigma^* &= \sqrt{\sigma^2 + M},
\end{aligned}$$

where M is the minimized value of $E\{(\beta_0 - \beta_0^* + (\beta_1 - \beta_1^*)Z + f(X) - \gamma^* X)^2\}$.

Suppose now that X, Z are independent, and assume (without loss of generality) that also $Ef(X) = 0$. Then the solution is

$$\begin{aligned}
\beta_0^* - \beta_0 &= 0 \\
\beta_1^* - \beta_1 &= 0 \\
\gamma^* &= E\{Xf(X)\} = Cov(X, f(X)) \\
\sigma^* &= \sqrt{\sigma^2 + E\{(f(X) - \gamma^* X)^2\}}.
\end{aligned}$$

Suppose now instead that $Cov(X, Z) \equiv E(XZ) = \rho$, but still $E(X) = E(Z) = E(f(X)) = 0$, while also assuming (without loss of generality) that $E(X^2) = E(Z^2) = 1$. Then we get

$$\begin{aligned}
\beta_0^* - \beta_0 &= 0 \\
\beta_1^* - \beta_1 &= \frac{E\{(Z - \rho X)f(X)\}}{1 - \rho^2} \\
\gamma^* &= \frac{E\{(X - \rho Z)f(X)\}}{1 - \rho^2}.
\end{aligned}$$

Note that $Cov(Z - \rho X, X) = 0$, so the numerator of the expression for $\beta_1^* - \beta_1$ is the expected value of a product of something that is uncorrelated with X times a function of X. Intuitively this should be a small number. In fact, if we further assume that (X, Z) is binormal, and that $f(X) = X^2 - E(X^2)$, then since $Var(X|Z) = 1 - \rho^2$ and $E(X|Z) = \rho Z$, we get

$$E(ZX^2) = E[ZE(X^2|Z)] = E[Z(Var(X|Z) + (E(X|Z))^2] = E[Z(1 - \rho^2 + \rho^2 Z^2] = 0,$$

which in fact implies that $\beta_1^* - \beta_1 = 0$ in this case.

Example - Weibull distribution

Now $\phi(x) = e^x e^{-e^x}$ and $\Phi(x) = 1 - e^{-e^x}$, so $\log \phi(x) = x - e^x$ and $\log(1 - \Phi(x) = -e^x$. Note also that $Ee^W = 1$. If we allow censoring, we get

$$\begin{aligned} E(D) &= E\left[\Delta \log \frac{\sigma^*}{\sigma} + \Delta W - 1 - \Delta\left(\frac{\sigma}{\sigma^*}W + \frac{\beta_0 - \beta_0^* + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)'\boldsymbol{Z} + f(X) - \gamma^* X}{\sigma^*}\right)\right. \\ &+ \left.\exp\left\{\frac{\sigma}{\sigma^*}W + \frac{\beta_0 - \beta_0^* + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)'\boldsymbol{Z} + f(X) - \gamma^* X}{\sigma^*}\right\}\right]. \end{aligned}$$

The last term equals (26) and is hence conditionally Weibull distributed given $\boldsymbol{Z}$ and X, with conditional expectation

$$\Gamma\left(1 + \frac{\sigma}{\sigma^*}\right) \exp\left\{\frac{\beta_0 - \beta_0^* + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)'\boldsymbol{Z} + f(X) - \gamma^* X}{\sigma^*}\right\}.$$

The above formula for $E(D)$ furthermore involves $E(\Delta) = P(\beta_0 + \boldsymbol{\beta_1}'\boldsymbol{Z} + f(X) + \sigma W < C)$ and also $E(\Delta W)$, which may be fairly complicated expressions.

Let us therefore below consider the non-censored case. Note here that $E(W) = -a$, where $a = 0.577215665\ldots$ is Euler's constant. The expression then becomes

$$\begin{aligned} E(D) &= \log \frac{\sigma^*}{\sigma} - a(1 - \frac{\sigma}{\sigma^*}) - 1 - \frac{\beta_0 - \beta_0^* + (\boldsymbol{\beta}_1 - \boldsymbol{\beta_1}^*)'E(\boldsymbol{Z}) + E(f(X)) - \gamma^* E(X)}{\sigma^*} \\ &+ \Gamma\left(1 + \frac{\sigma}{\sigma^*}\right) E\left[\exp\left\{\frac{\beta_0 - \beta_0^* + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)'\boldsymbol{Z} + f(X) - \gamma^* X}{\sigma^*}\right\}.\right] \end{aligned}$$

Now if we assume that $X, \boldsymbol{Z}$ are independent, with $\boldsymbol{Z}$ being multinormally distributed with covariance matrix given by the identity matrix, while $E(X) = E(f(X)) = 0$ (but X not necessarily normal), then $E(D)$ becomes

$$\log \frac{\sigma^*}{\sigma} - a(1 - \frac{\sigma}{\sigma^*}) - 1 - \frac{\beta_0 - \beta_0^*}{\sigma^*} + \Gamma\left(1 + \frac{\sigma}{\sigma^*}\right) \exp\left\{\frac{\beta_0 - \beta_0^*}{\sigma^*} + \frac{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)}{2\sigma^{*2}}\right\}$$

$$\times E \exp\left\{\frac{f(X) - \gamma^* X}{\sigma^*}\right\}.$$

It is clear from this expression that the solution for $\boldsymbol{\beta}_1^*$ is $\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1 = 0$, but note that multinormality of $\boldsymbol{Z}$ is crucial for this result. No explicit solution can be found for the other parameters, however, so numerical methods are needed. But it can be seen that if (i) X has a distribution that is symmetric around 0, i.e. X and $-X$ has the same distribution, and (ii) $f(X) = g(X) - E(g(X))$ where $g(-x) = g(x)$, then $\gamma^* = 0$. To see this, differentiate the above expression with respect to γ^* and check that the result equals 0 if γ^* is set to 0.

Example - using simulated data for Weibull distribution

Aaserud (2011) gives an example to show how the starred parameters can be found by simulation. It is clear that the precision of the obtained values can in principle be made as small as desired by increasing the number of simulations.

Aaserud (2011) considered the true Weibull regression model

$$\log T = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + X^2 + \sigma W$$

with $\beta_0 = 0,\ \beta_1 = 5,\ \beta_2 = 0.2,\ f(x) = x^2,\ \sigma = 2$; W being Gumbel distributed, independent of Z_1, Z_2, X, which were assumed independent and standard normally distributed. 1,000,000 non-censored observations $(t_i, z_{i1}, z_{i2}, x_i)$ were then drawn from the true model, while the following misspecified model was fitted by maximum likelihood,

$$\log T = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \gamma X + \sigma W$$

By White (1982) it follows that the estimates of the parameters are approximately equal to the starred parameters. The following values were obtained, $\hat{\beta}_0 = 1.2479,\ \hat{\beta}_1 = 5.0060,\ \hat{\beta}_2 = 0.2169,\ \hat{\gamma} = 0.0099,\ \hat{\sigma} = 3.14$.

From the theoretical computations in the Weibull example considered above, it follows that the true values of the starred parameters are $\beta_1^* = \beta_1 = 5$, $\beta_2^* = \beta_2 = 0.2$ and $\gamma^* = 0$. The theoretical results are thus confirmed by the simulation. In the analytical approach we did not get simple expressions for the remaining parameters, β_0 and σ, and it is seen from the simulation results that these are changed in the misspecified model. The reason is that the effect of the X^2 term has to be assimilated in the constant term β_0 and σW.

C The covariate order method for censored exponential regression

Exponential regression means to estimate the hazard rate $\lambda(\boldsymbol{X})$ as a function of the covariates $\boldsymbol{X}$ for exponentially distributed data. A possible way of doing this is using the so called *covariate order method*, described in more detail in (Kvaløy and Lindqvist, 2003, 2004).

In the case of a single covariate X, the basic idea of the method is to arrange the data in increasing order of X, and then define a certain point process based on the corresponding event data. We start by presenting the basic method, indicating how testing procedures follow from the same idea, and then show how all this can be applied to Cox-Snell residuals in AFT models.

Thus, assume that we have n independent observations $(T_1, \delta_1, X_1), \ldots, (T_n, \delta_n, X_n)$, where $T_i = \min(Y_i, C_i)$, and Y_i given $X_i = x$ is exponentially distributed with hazard rate $\lambda(x)$. The method starts by arranging the observations $(T_1, \delta_1, X_1), \ldots, (T_n, \delta_n, X_n)$ such that $X_1 \leq X_2 \leq \cdots \leq X_n$. Next, for convenience, divide the observation times by the number of observations, n. Then let the scaled observation times $T_1/n, \ldots, T_n/n$, irrespectively if they are censored or not, be subsequent intervals of an artificial point process on a "time" axis s. For this process, let points which are endpoints of intervals corresponding to *non-censored* observations be considered as events, occurring at times denoted $S_1, \ldots, S_r$ where $r = \sum_{j=1}^{n} \delta_j$. This is visualised in Figure 6, for an example where the ordered observations are $(T_1, \delta_1 = 1), (T_2, \delta_2 = 0), (T_3, \delta_3 = 1), \ldots, (T_{n-1}, \delta_{n-1} = 0), (T_n, \delta_n = 1)$.

First notice that if there is no covariate effect, i.e. $\lambda(x) = \lambda$, then the process $S_1, \ldots, S_r$ is a homogeneous Poisson process. The test presented in Section C.1 is based on this observation. Further, if $\lambda(x)$ is reasonably smooth and not varying too much, then the process

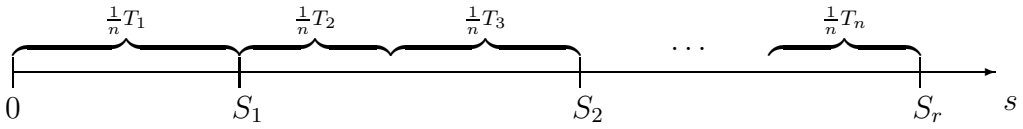

Figure 6: Construction of the process $S_1, \ldots, S_r$.

$S_1, \ldots, S_r$ could be imagined to be nearly a nonhomogeneous Poisson process for which the intensity can be estimated by for instance kernel density estimation. Notice that the true conditional intensity of the process $S_1, \ldots, S_r$ at a point w is $n\lambda(X_I)$ where I is defined from $\sum_{i=1}^{I-1} T_i/n < w \leq \sum_{i=1}^{I} T_i/n$. Thus from the estimated intensity of the process, say $\hat{\rho}_n(w)$, an estimate of $\lambda(X_I)$ is found as $\hat{\lambda}(X_I) = \hat{\rho}_n(w)/n$.

The relationship between covariate values and corresponding points in the process $S_1, \ldots, S_r$ can generally be defined for instance by the simple function

$$\tilde{s}(x) = \frac{1}{n} \sum_{i=1}^{j} T_i, \quad X_j \leq x < X_{j+1}, \tag{34}$$

and the estimator can then be written $\hat{\lambda}(x) = \hat{\rho}(\tilde{s}(x))/n$.

A number of different estimators $\hat{\rho}(\cdot)$ can be obtained, one simple approach is to use a kernel estimator giving

$$\hat{\lambda}(x) = \frac{1}{nh_s} \sum_{i=1}^{r} K\left(\frac{\tilde{s}(x) - S_i}{h_s}\right) \tag{35}$$

Here $K(\cdot)$ is a positive kernel function which vanishes outside [-1,1] and has integral 1, and h_s is a smoothing parameter. Under certain mild regularity conditions it can be shown that this is a uniformly consistent estimator of $\lambda(x)$, see Kvaløy and Lindqvist (2004) for proofs and further details. The value of the smoothing parameter can for instance be chosen using a likelihood cross-validation criterion. To avoid the estimate $\hat{\lambda}(x)$ to be seriously downward biased near the endpoints, the reflection method for handling boundary problems in density estimation is used, see for example Silverman (1986).

It may be remarked that for the situation with several covariates, the above method can be used to fit generalized linear models in an iterative manner (Kvaløy and Lindqvist, 2004). Extensions of the method to Cox-regression is relatively straightforward by using appropriate time transformations, see e.g. Kvaløy and Lindqvist (2003), but this is not the focus in the present paper.

C.1 Testing for covariate effect

Recall that if there is no covariate effect, that is $\lambda(x) \equiv \lambda$, then the process $S_1, \ldots, S_r$ is a homogeneous Poisson process (HPP). This observation suggests that in principle any statistical test for the null hypothesis of an HPP versus various non-HPP alternatives can be applied to test for covariate effect in exponential regression models.

A detailed account of this approach for testing for covariate effect in event time data is given in Kvaløy (2002), who studied a number of different tests constructed based on the covariate order method. The recommendation is to use an Anderson-Darling type test which turns out to have very good power properties against both monotonic and non-monotonic alternatives to constant $\lambda(x)$, and thus is a good omnibus test for covariate effect which can be used in any event time model. The test can be used both for continuous and discrete covariates.