GOODNESS OF FIT OF RELATIVE SURVIVAL MODELS

Janez Stare
Department of Medical Informatics
University of Ljubljana,
Vrazov trg 2, SI-1000 Ljubljana; Slovenia
e-mail: janez.stare@mf.uni-lj.si

Maja Pohar
Department of Medical Informatics
University of Ljubljana, Slovenia

Robin Henderson
Mathematics and Statistics
University of Lancaster, UK

November 13, 2008

Abstract

Additive regression models are preferred over multiplicative models in analysis of relative survival data. Such preferences are mainly grounded in practical experience with mostly cancer registries data, where the basic assumption of the additivity of hazards is more likely to be met. Also, the interpretation of coefficients is more meaningful in additive than in multiplicative models. Nonetheless, the question of goodness of fit of the assumed model must still be addressed, and while there is an abundance of methods to check the goodness of fit of multiplicative models, the respective arsenal for additive models is almost empty. We propose here a variety of procedures for testing the null hypothesis of a good fit. These are based on partial residuals defined similarly to Schoenfeld residuals familiar for Cox model diagnostics. The tests have appropriate sizes under the null hypothesis, and good power under different alternatives. We investigate their performance through simulations and apply the methods to data from a study into survival of colon cancer patients.

1 Introduction

Relative survival methods are used when observed survival experience (O) in a cohort of patients is to be compared with expected or population (P) survival obtained from life tables. Such methods are useful for studies that aim to estimate cause-specific mortality, but lack information on individual causes of death.

The cumulative relative survival function is defined (eg Ederer et al, 1961) as

$$r(t) = \frac{S_O(t)}{S_P(t)}$$

where $S_O(t)$ denotes observed survival and $S_P(t)$ stands for population or expected survival. If $r(t)$ is a decreasing function then an additive hazard relationship is implied, namely

$$\lambda_O(t) = \lambda_P(t) + \lambda_E(t)$$

where λ_E is the excess hazard experienced by the cohort. This relationship forms the basis of the additive regression model for the effects of $p-$dimensional covariates X on relative survival. We will consider here the commonly used model

$$\lambda_O(t, X) = \lambda_P(t, X_1) + \lambda_E^{(0)}(t)e^{\beta X} \tag{1}$$

where β stands for a $p-$dimensional vector of coefficients. We write $X = (X_1, X_2)$ with X_1 denoting the variables by which population values are stratified, typically age, sex, calendar year and perhaps ethnicity, and $\lambda_E^{(0)}(t)$ is the baseline excess hazard. Of course, not all X_1 variables need be included in the excess hazard. Covariates can vary over time, but we don't make that explicit in the notation. For estimation, the baseline excess is often taken to be piecewise constant over a partition of the follow-up interval $(0, \tau)$, but more flexible methods exist (see Sasieni 1996 and Giorgi et al 2003).

The model (1) has had considerable attention and success (see eg Hakulinen and Tenkanen 1987, Estève et al 1990, Dickman et al 2004, and references therein), so much so that the phrase "relative survival model" is sometimes used specifically for this additive hazard class. In a broader sense the term means any model which relates observed survival to background, including also the multiplicative model class (Buckley 1984, Andersen et al 1985, Andersen and Væth 1989) and the transformation approach (Stare et al, 2005).

The purpose of this paper is to consider goodness-of-fit methods for model (1). Standard diagnostics for Cox proportional hazards models apply directly under the transformation approach and can be adapted easily for multiplicative models, but there has been little development of diagnostic tools specifically for (1). Giorgi et el (2003) apply the method of Abrahamowicz et al (1996) to the additive model to relax the proportionality assumption and thereby improve the fit, and Bolard et al (2001) use a piecewise proportional hazard model for the same purpose. For grouped data, generalised linear model diagnostics are available (Dickman et al, 2004), although to our knowledge their properties for relative survival have not been investigated. We will concentrate on methods for non-grouped data which are aimed at detecting time-varying effects of covariates on the excess risk, which can be described by

$$\lambda_O(t, X) = \lambda_P(t, X_1) + \lambda_E^{(0)}(t) e^{\beta(t)X}, \tag{2}$$

for unspecified functions $\beta(t)$.

In Section 2 we introduce a form of residual for the additive model, akin to Schoenfeld (1982) partial residuals for the proportional hazards model. These contrast the covariates of subjects dying at any given time with the corresponding model-based expectation given the risk sets. Plots of these residuals can be used to provide informal guidance as to the possibility of time-varying effects. In Section 3 we investigate three proposed test statistics based on cumulative sums of the scaled partial residuals. Simulations are used to validate performance and compare power in Section 4. In Section 5 we illustrate the methods using data on colon carcinoma, supplied by the Finnish Cancer Registry (Dickman et al 1999). Some closing remarks in Section 6 complete the paper.

2 Partial residuals for the additive model

Schoenfeld's (1982) partial residuals are a useful diagnostic tool for the Cox model, where they naturally arise from the estimating equations obtained from partial likelihood. The fact that they are components of the score function enables us to deduce some useful properties. Nothing similar follows from the score function for the additive model, but we nevertheless suggest defining Schoenfeld-like residuals in an analogous way. Denoting by X_i the covariate vector of the subject who fails at t_i, we define the residuals via

$$U_i(\beta) := X_i - \frac{\sum\limits_{j \in R_i} X_j \left\{ \lambda_P(t_i, X_{1j}) + \lambda_E^{(0)}(t) e^{\beta X_j} \right\}}{\sum\limits_{j \in R_i} \left\{ \lambda_P(t_i, X_{1j}) + \lambda_E^{(0)}(t) e^{\beta X_j} \right\}}.$$

Here X_j is the vector of covariates for person j at event time t_i and $R_i = R(t_i)$ is the risk set at time t_i. We drop the argument t_i in some equations for convenience, and will also assume that the data are ordered by failure time, so $t_i \leq t_{i+1}$. Note that to calculate these residuals, $\lambda_E^{(0)}(t)$ will have to be estimated. It always is, but as mentioned before, methods to do this differ, and the estimates will always depend on the assumed model.

We will write, for clarity, $\beta_0(t)$ for the true parameter value. Then

$$\begin{aligned} \frac{\left\{ \lambda_P(t_i, X_{1j}) + \lambda_E^{(0)}(t) e^{\beta_0(t_i) X_j} \right\}}{\sum\limits_{k \in R_i} \left\{ \lambda_P(t_i, X_{1k}) + \lambda_E^{(0)}(t) e^{\beta_0(t_i) X_k} \right\}} &= \lim_{\Delta t \to 0} \frac{P(T_j \in [t_i, t_i + \Delta t) | T_j \geq t_i)}{\sum\limits_{k \in R_i} P(T_k \in [t_i, t_i + \Delta t) | T_k \geq t_i)} \\ &= \lim_{\Delta t \to 0} \frac{P(T_j \in [t_i, t_i + \Delta t) | T_j \geq t_i)}{P(\bigcup\limits_{k \in R_i} (T_k \in [t_i, t_i + \Delta t) | T_k \geq t_i))} \end{aligned}$$

is the probability of the individual j dying at time t_i, conditional on exactly one event occurring at time t_i in the risk set R_i. The expected value of $U_i(\beta(t_i))$, considered as a function of a random covariate, is thus zero when we have the true hazard model and this motivates us to write

$$U_i(\beta) = Z_i - \hat{E}_\beta(Z|R_i),$$

as in the Cox model. Note that U_i here does not refer to components of the score function, we simply use the same notation to stress the analogy. In notation we also stress that the expectations are with respect to the model and depend on β. Properties of these residuals are described in the Appendix. In particular $U_i(\beta)$ and $U_j(\beta)$ are uncorrelated for $j \neq i$, and

$$E[U(\beta)|R_i] \simeq \frac{\partial}{\partial\beta}\{\hat{E}_\beta(Z|R_i)\}\left(\beta_0(t_i) - \beta\right).$$

Hence

$$\beta_0(t_i) \simeq \beta + \left(\frac{\partial}{\partial\beta}\{\hat{E}_\beta(Z|R_i)\}\right)^{-1} E[U(\beta)|R_i], \tag{3}$$

and we can obtain information about the behaviour of $\beta_0(t)$ from a smooth trace through the scaled residuals. Which β is plugged in on the right side of (3) does not really matter, but one needs to keep in mind that the neglected part in the above Taylor series expansion might depend on it. In practice, one would most often estimate β from a model with assumed constant effects. In addition, equation (3) can also be used to graphically estimate the adequacy of our estimation of $\beta(t)$, like the one using splines for example. In that case, we need to plot the difference $\beta_0(t_i) - \beta(t_i)$ (meaning the second term in equation(3)), which should be more or less constant, if our estimates are good.

3 Goodness-of-fit statistics

An expression for the variance $V_i(\beta)$ of the residual $U_i(\beta)$ is given in the Appendix. With this we form standardized residuals $R_i(\beta) = U_i(\beta)/\sqrt{V_i(\beta)}$.

In turn, working from the ideas of O'Quigley (2003) and O'Quigley and Stare (2003), we form a cumulative sum process:

$$B_n(\beta, \frac{k}{n}) := \frac{1}{\sqrt{n}} \sum_{i=1}^{k} R_i(\beta),\ k = 1, \ldots, n;\ B(\beta, 0) := 0$$

Here n denotes the number of events, not the sample size. This process is only defined on n equispaced points k/n of the interval $[0, 1]$, but we can extend the definition to the whole interval using a linear interpolation so that, for $u \in (k/n, (k+1)/n)$, we write:

$$B_n^{(c)}(\beta, u) = B_n(\beta, \frac{k}{n}) + (un - k)(B_n(\beta, \frac{k+1}{n}) - B_n(\beta, \frac{k}{n})). \qquad (4)$$

Since the standardized residuals are uncorrelated, the Central Limit Theorem shows that $B_n^{(c)}(\beta_0, u)$ converges in distribution to a zero-mean Gaussian random variable $B^{(c)}(\beta, u)$ with variance u. Also the asymptotic covariance between $B_n^{(c)}(\beta_0, u)$ and $B_n^{(c)}(\beta_0, u + s)$ is just u. These are the marginal and moment properties of Brownian motion, leading us to construct a bridge process

$$BP(\beta, u) = B^{(c)}(\beta, u) - uB^{(c)}(\beta, 1),$$

with sample version

$$BP_n(\beta, \frac{k}{n}) = \frac{1}{\sqrt{n}} \left\{ \sum_{i=1}^{k} R_i(\beta) - \frac{k}{n} \sum_{i=1}^{n} R_i(\beta) \right\},$$

so that we can use standard results from Brownian bridge theory (Resnick, 1992) to approximate the distribution of various test statistics.

The null hypothesis

Before we describe the proposed tests we want to stress that the null hypothesis one would in general want to test is that the assumed model generated the data. The standard additive model has two assumptions: about the form of the excess hazard and about the additivity. Either can be at fault and 'guilty' for the bad fit. While one can check the fit of any model, for example one with given functional form for $\beta(t)$, it will most often be the assumption of proportional excess hazards (or constant effects) that will be under

scrutiny. So to ease the narrative we will in what follows have such a null hypothesis in mind. To be specific, $H_0 : \beta = constant$. This is of course conditional on other assumptions of the model (additivity, form of the baseline excess hazard) being correct.

That such a null hypothesis will often be unrealistic is of course to be expected, and suggestions as to how to deal with non-proportionality have appeared in the literature. Giorgi et al 2003 used B-splines to model changing effects. These are flexible methods and will often give satisfactory results, but they do not guarantee that the final model fits well, and methods to check the fit are still needed. We give an example in section 4.

Maximum bridge value

Under the null, a Brownian bridge process is more or less close to zero on the whole interval. But, if β is not constant in time, we expect to see this trend reflected in the motion of $BP(\beta, u)$. A sensible test statistic is therefore

$$T_1 = \max_k |BP_n(\beta, k/n)|$$

The distribution of the maximum absolute value of a Brownian bridge $BB(u)$ is

$$P(\max_{u\in[0,1]}(|BB(u)|) \leq x) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2x^2}, \quad x > 0. \tag{5}$$

We will use this to approximate the null distribution of the test statistic. Accuracy of the approximation is explored in Section **??**.

Maximum bridge value, constructed from weighted residuals

So far, the residuals are treated equally after standardisation. Instead we may prefer to attach more importance to some regions of the time scale than others, for instance early times where risk set sizes are large. This can be achieved by introducing non-negative weights w_i, $i = 1, ..., n$, with $\sum_{i=1}^{n} w_i = 1$ but otherwise arbitrarily defined, and measuring time in a transformed scale $u_k = \sum_{i=1}^{k} w_i$.

The cumulative weighted sum

$$B_n^w(\beta, u_k) = \sum_{i=1}^{k} R_i(\beta)\sqrt{w_i},$$

can be interpolated in a similar way as (4) and for $u \in [0, 1]$ converges to a Gaussian process on $B^w(\beta, u)$, with

$$\text{var}(B^w(\beta, u)) = u, \ \ \text{cov}(B^w(\beta, u), B^w(\beta, u + s)) = u.$$

Hence the maximum absolute value of the new sample bridge

$$BP_n^w(\beta, u_k) = \sum_{i=1}^{k} R_i(\beta)\sqrt{w_i} - u_k \sum_{i=1}^{n} R_i(\beta)\sqrt{w_i}$$

can be used as test statistic, and we can approximate its distribution under the null hypothesis using (5). Of course if all the weights are equal ($w_i = 1/n$ for each i), the two processes $BP_n^w(\beta, u)$ and $BP_n(\beta, u)$ are equivalent.

In order to give more weight to the events at the beginning of our study, we can set the weights to be proportional to the number of patients at risk at each time of event. We will use T_2 to denote the test statistic with this choice of weights.

Cramér - Von Mises statistic

The Cramér - Von Mises statistic is defined as

$$v^2 := \int_0^1 BB^2(t)dt - (\int_0^1 BB(t)dt)^2$$

and distributed as (Csörgő et al, 1996)

$$P(v^2 \leq x) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2\pi^2 x}, x > 0. \tag{6}$$

This statistic exploits the whole path of a bridge process rather than the maximum only. Our third test statistic is the associated sample version

$$T_3 = \frac{1}{n}\sum_{k=1}^{n} BP_n^2(\beta, \frac{k}{n}) - \left(\frac{1}{n}\sum_{k=1}^{n} BP_n(\beta, \frac{k}{n})\right)^2 .$$

4 Simulations

In this section we investigate the proposed methods via simulations, and in the next we apply them to a real data set. We designed our simulations to reasonably reflect the type of data we will be interested in practice. The baseline excess hazard was taken to be piecewise constant over a partition $0 = t_0 < t_1 < t_2 < \ldots < t_K = \tau$ of the follow-up interval $(0, \tau)$, so that the generating model was

$$\lambda_O(t, X) = \lambda_P(t, X_1) + e^{\sum_k \alpha_k I_k(t)} e^{\beta X}, \tag{7}$$

where $I_k(t)$ is an indicator for $t \in (t_{k-1}, t_k]$.

The population hazard was taken from Slovenian national life tables, subdivided by age, sex, and year of birth. Covariates X included in the regression model (7) were age, sex, and year of diagnosis. Ages were drawn from a uniform distribution between 50 and 80, males and females were equally likely, and we took two diagnosis years, 1980 and 1990, again equally likely. Maximum follow-up time τ was set so that in each scenario 90% of subjects were expected to die in the follow-up period, giving 10% Type 1 censoring. In addition, in some simulations we added a further 40% random exponential censoring. The partition $0 = t_0 < t_1 < t_2 < \ldots < t_K = \tau$ had annual intervals for the first five years, longer periods thereafter, depending upon the value of τ. The baseline excess parameters were all set to be the same in generating data, ie $\alpha_1 = \ldots = \alpha_K = \alpha$, say, though this was not assumed in estimation. We used two values of α, -1.5 and -3. The three covariates age,

α	n	censoring	T_1	T_2	T_3
-1.5	250	0.1	0.041	0.037	0.041
		0.5	0.041	0.034	0.056
	500	0.1	0.050	0.042	0.051
		0.5	0.045	0.030	0.050
	1000	0.1	0.050	0.040	0.061
		0.5	0.051	0.039	0.044
-3	250	0.1	0.042	0.037	0.042
		0.5	0.048	0.047	0.054
	500	0.1	0.034	0.033	0.040
		0.5	0.036	0.029	0.040
	1000	0.1	0.041	0.040	0.031
		0.5	0.048	0.045	0.056

Table 1: *The proportion of tests rejecting under the null hypothesis (*1000 *simulations).*

sex and year of diagnosis were always included in the model for estimation, but the true regression coefficient β was non-zero only for the sex effect. For $\beta_{sex} = 1$, the value of -1.5 for α results in roughly 80% of the deaths due to excess risk (and 20% due to population risk), while -3 gives 50:50 odds.

We first examined how empirical distributions of our statistics approximate theoretical ones. Figure 1 shows empirical distributions of the T_1 and T_3 statistics under the null and three different alternative hypotheses. The top left graph shows an almost perfect match between the theoretical and empirical distributions under the null hypothesis. Graphs do not include T_2 as the curves were almost identical to T_1 for this particular simulation.

Table 1 shows achieved test sizes for nominal 5% tests, based on the asymptotic distributions, when the null hypothesis is true. The average rejection rate for T_1 was 0.044, for T_2 0.038, and for T_3 0.047, and the rejection rates do not seem to depend on the sample size, censoring proportion or the value of α. We find these results very satisfactory.

We then considered performance under a variety of models which allowed the sex coefficient to change over follow-up time. Results for two are presented

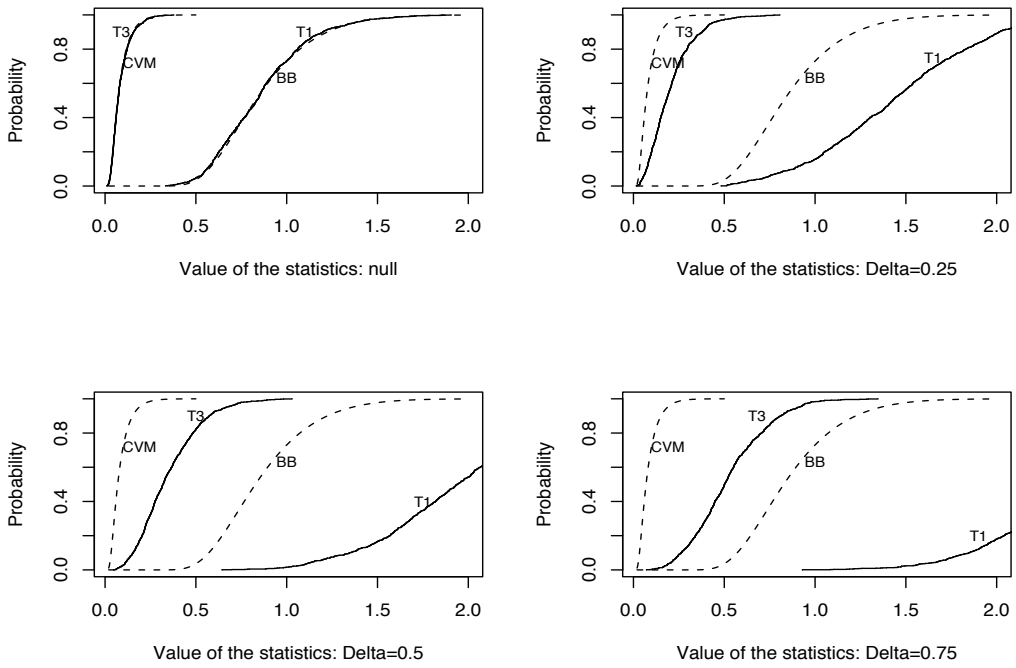

Figure 1: *Comparison of empirical distributions (solid curves) of test statistics* T_1 *and* T_3 *with corresponding asymptotic distributions BB and CVM from (5) and (6)(dotted curves). The first graph shows the distributions under the null (*1000 *simulations, sample size* $n = 500$*, censoring*=10%*,* $\alpha = -1.5$*), in the next ones, the parameter* β *for sex is simulated to change from 1 to* $1-\Delta$ *for* $\Delta = 0.25, 0.5, 0.75$ *at proportion* $\theta = 0.5$*.*

here. The first assumed a simple step change in sex effect of size Δ at a follow-up time, chosen to give proportion θ of events before the step. Distribution functions of T_1 and T_3 are given in Figure 1 and show clear movement away from the null and hence the potential for good power. Some illustrative plots of smoothed scaled residuals (see formula 3) are in Figure 2. In R the smoothing is done using splines, and we used five knots in this case. Simulation results are in Table 2 and Table 3. As expected, higher censoring lowers rejection rates and a higher change in β increases them. For this scenario T_2 outperforms T_1 when the change-point is early, the opposite when

the change-point θ is later. Both statistics have better power than T3 for this form of alternative. Table 3 illustrates the effect of the sample size and the change in β for T_1. It should be noted that a change of $\Delta = 0.75$ will be detected only half of the time with sample sizes of 250. We also compared our methods to the usage of B-splines as described in Giorgi et al, for the two examples described on page 2773 of that paper. In the first example, β for a binary variable x changes gradually from -0.5 to 1.2 (as read from the Figure 1 of their paper) and in the second example β starts at 0.5, goes down to -1.2 and then up again all the way to 1.5 (see Giorgi et al for details). These are very big changes, and they describe crossing hazards in both cases. One would expect such departures from the proportional excess hazards assumption to be detected almost always with any decent sample size. That this really happens is shown in Table 4. What is interesting to us is that T_1 is at least as good as the spline method in both examples, and T_3 in the second example, even though both examples are tailored to the usage of splines. For the second example estimation problems for splines were reported in 3% of the cases (we used the program written by Giorgi to fit splines). We do not of course over-interpret this specific comparison, but it does indicate that our tests can be competitive to the Giorgi et al procedure even in examples where the spline approach might a priori be expected to have advantages. Further investigations might be interesting, but the aim of this paper is to introduce a simple, theoretically sound approach to checking goodness of fit of relative survival models.

To illustrate a situation where T_3 might be preferable we simulated data where the coefficient for sex alternated between 1 and 0, and the changes occurred after 10%, 50%, and 90% of the failures. For the sample size of $n = 500$, T_3 rejected the null in 91% of the samples, as compared to 68% and 71% for T_1 and T_2. The second example of Giorgi et al, mentioned above, is also better suited for T_3 than T_1, although this is not obvious due to big changes in β.

Other simulations, not reported, showed qualitatively similar results: the Brownian bridge approximations for the distributions of T_1, T_2 and T_3 are good for samples of about 200 or more, and power has the expected properties.

We conclude this section with an example in which we show that simply

applying splines does not guarantee a good fit, and that the fit should still be checked. We simulated a data set of size 300, with parameters like in the first example of Giorgi et al described above, except that the change in parameter from -0.5 to 1.2 occurred at midpoint. Fitting B-splines of Giorgi et al gave a highly significant result ($\chi^2 = 21.79, df = 5, p = 0.0006$), and T_1 was even more significant ($T_1 = 3.71, p = 2.34 \cdot 10^{-12}$). If we now check the model fitted with splines, T_1 is still significant ($T_1 = 1.71, p = 0.0058$). Only after we fit the change point model, is T_1 satisfied ($T_1 = 0.768, p = 0.597$). The respective Brownian bridge processes are plotted in Figure 3.

We are well aware of course that such models might be unrealistic, we use them for the purpose of illustration. Splines will work well most of the time, and even in the above example one could do better if values other than default were used. The point we are making is that it is one thing to detect a bad fit, and the other to fix it. Sometimes these concepts get mixed up.

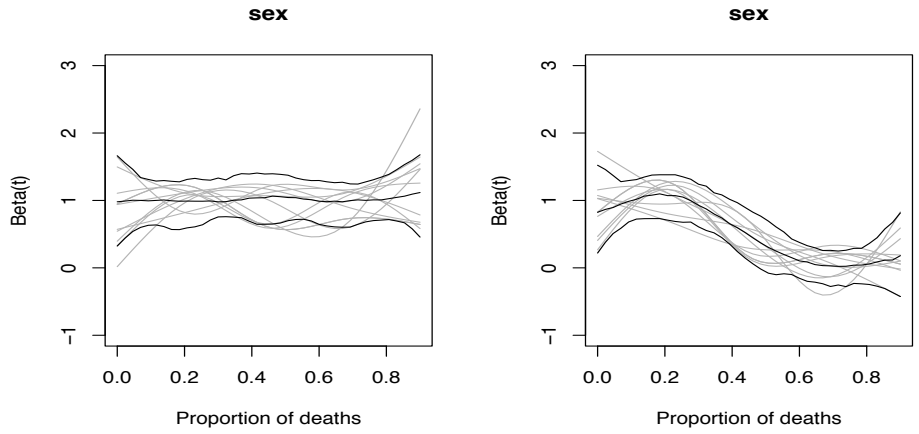

Figure 2: *Behaviour of β for sex in time (obtained from smoothing the scaled residuals, see formula 3), where time is presented as proportion of deaths. On the left figure, the data are simulated with a constant $\beta = 1$, on the right, β changes to* 0 *at proportion θ=0.5. The grey lines are ten examples, the black lines represent the 5th, 50th and 95th quantile. (*100 *simulations, sample size* $n = 500$, *censoring* $= 10\%$, $\alpha = -1.5$*).*

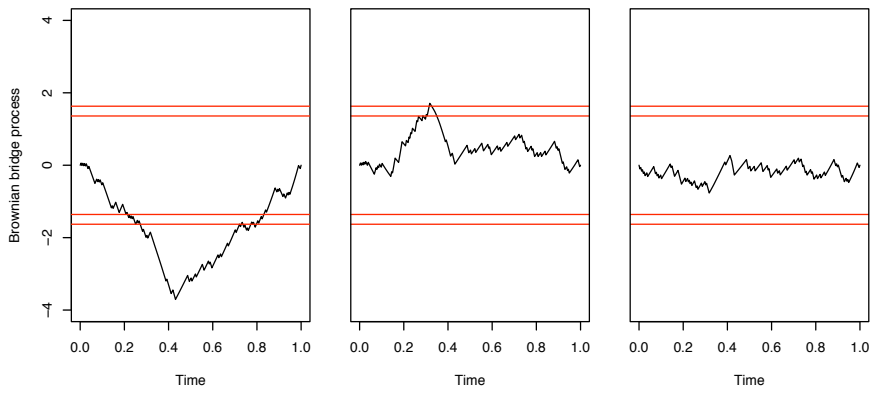

Figure 3: *Brownian bridge processes for the example of Giorgi et al (see text). From left to right: fit with constant β, fit with B-splines, fit with a change point.*

Δ	θ	censored	T_1	T_2	T_3
0.5	0.25	0.1	0.302	0.417	0.263
		0.5	0.171	0.233	0.166
	0.5	0.1	0.569	0.504	0.481
		0.5	0.383	0.252	0.328
0.75	0.25	0.1	0.685	0.806	0.591
		0.5	0.399	0.535	0.342
	0.5	0.1	0.898	0.862	0.833
		0.5	0.718	0.610	0.642

Table 2: *The proportion of tests rejecting, depending on the size of Δ and θ and the amount of censoring (*1000 *simulations, $\alpha = -1.5$, sample size = 500).*

5 Application

To illustrate the methods on a real data set we chose data on localized colon carcinoma, supplied by the Finnish Cancer Registry. The data set consists of

n	$\Delta = 0.25$	$\Delta = 0.5$	$\Delta = 0.75$	$\Delta = 1$
250	0.084	0.284	0.587	0.819
500	0.154	0.569	0.898	0.990
1000	0.349	0.887	0.996	1.000

Table 3: *The proportion of simulations in which test T_1 rejects the null hypothesis, depending on sample size n and the size of parameter change Δ (*1000 *simulations, $\theta = 0.5$, censoring*= 10%*).*

Example	T_1	LRT	T_3
a	0.99	0.975	0.885
b	1	1	1

Table 4: *The proportion of simulations in which tests T_1, likelihood ratio for splines, and T_3 reject the null hypothesis of proportional excess hazards for the two examples of Giorgi et al (200 simulations, n =* 300*, proportion censored after the last failure* 13% *and* 12%*). See text for details.*

6274 patients diagnosed during 1975–1994 with follow-up to the end of 1995 (Dickman et al 1999). It provides a typical example of non-proportional excess hazards with respect to age, as discussed by Dickman et al (2004). In the analysis we were interested in the first five years of follow up, in which 2247 deaths (36%) occurred, and 19% were censored. The remaining 45% survived longer than 5 years. We fitted a model using covariates age (taking values between 18 and 99), sex (1 = male, 2 = female), and diagnosis year (grouped into 1 = 1975-1984 and 2 = 1985-1994) and 5 yearly follow-up intervals. Figure 4 gives smoothed residuals for sex and age, and Figure 5 the respective plots of bridge processes. The plots for age illustrate how looking only at the smoothed residual plots may not be enough to decide whether the fit is good, since the choice of the scale affects the impression. We see that the effect is declining at first and becomes constant sometime before 2 years, but a formal test is necessary. Figure 5 provides very strong evidence that the age effect is not time-constant, whereas the sex effect might be. The p-values for all three proposed tests are very small (10^{-13}), and none are significant for sex (respective p-values were 0.27, 0.36, and 0.36). To investigate further we re-fitted the model with the age coefficient allowed to change over time,

guided by Figure 5. Results for this and the original model are summarized in Table 5. Higher age leads to high excess hazard for about a year, but after that age is less important. The bridge processes for the new fit are given in Figure 6 and we can see that while the process for sex remained practically unchanged, the process for age now lies within the limits indicating a much better fit. We also tried fitting a model allowing sex to change over time, but as expected from Figure 5, the coefficients remained insignificant in all the intervals.

	variable	coef	se	z
model 1	age	0.013	0.003	4.016
	sex	−0.053	0.077	−0.688
	year	−0.317	0.074	−4.258
model 2	age$[0, 0.25)$	0.061	0.006	10.750
	age$[0.25, 0.5)$	0.035	0.007	5.194
	age$[0.5, 1)$	0.040	0.006	6.527
	age$[1, 2)$	0.009	0.006	1.402
	age$[2, 5)$	−0.009	0.005	−1.713
	sex	−0.098	0.073	−1.337
	year	−0.294	0.072	−4.113

Table 5: *The results of the fit using a constant age effect (model 1) and the one allowing for changing coefficients in pre-set time intervals (model 2). Five one-year long follow-up intervals were used in both cases.*

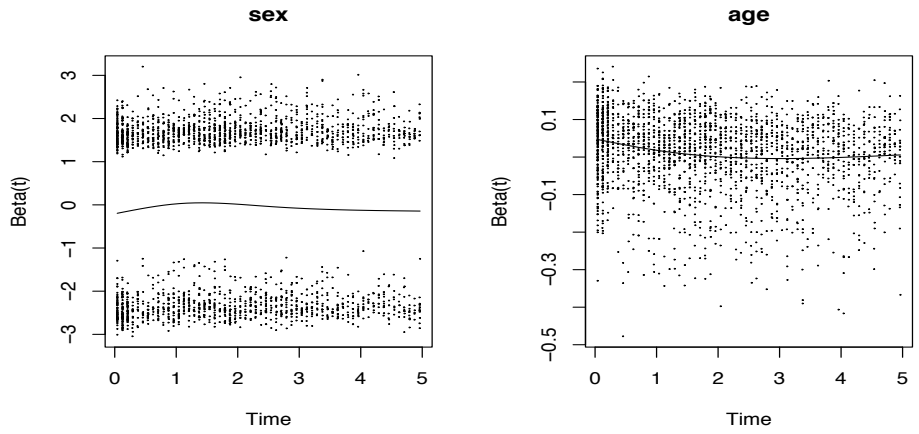

Figure 4: *Plots of smoothed residuals for sex and age (colon cancer data). Time scale is in years.*

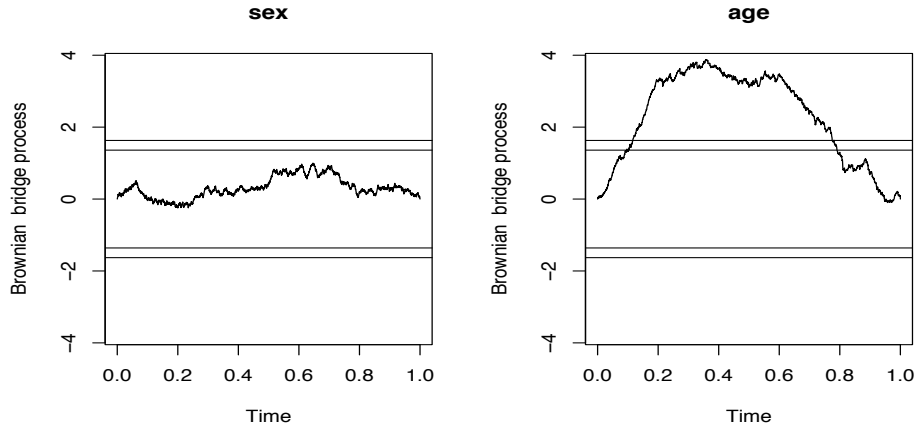

Figure 5: *Brownian bridge process for sex and age of the colon cancer data. The process changes at equidistant points of the standardized time scale. The horizontal lines represent the* 95% *and* 99% *confidence intervals for the maximum absolute value of a Brownian bridge, i.e. the* T_1 *statistic.*

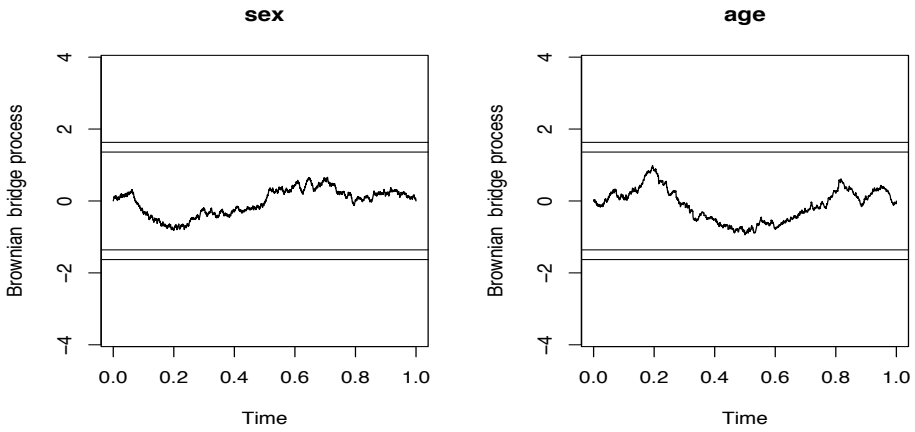

Figure 6: *The bridge processes for age and sex corresponding to the fit of model 2, given in Table 5. The horizontal lines represent the* 95% *and* 99% *confidence intervals for the maximum absolute value of a Brownian bridge, i.e. the* T_1 *statistic.*

6 Discussion

We presented a graphical method for checking the assumptions of the functional form for parameters (usually assumed to be constant), and three methods for testing these assumptions, based on the empirical random processes. The main idea behind these methods is the definition of residuals, akin to Schoenfeld residuals known for the Cox model diagnostics. A simple smoothed graph of these residuals seems to tell an informative story about the behaviour of the coefficients, apart from testing the null hypothesis. Nevertheless, as the example in the section 5 suggests, it is always good to support graphical procedures by formal test. For this, we introduced three tests, tailored for different departures from the null hypothesis. The first, based on the maximum value of the bridge process, is expected to work well when the estimated (fixed) β is really some kind of an average of otherwise increasing or decreasing values. The weighted version will obviously perform better when we want to give more weight to certain regions of the time scale. The Cramér Von Mises statistic performs better when β fluctuates around some average value. But the main purpose of this paper was to show the validity of the constructed tests, and not so much the exploration of their performance under different alternatives.

We provided some comparison with the approach of Giorgi et al (2003), mainly to point out the difference between modelling time-dependent effects, and checking goodness of fit of a model. For modelling, our graphical method can be seen as an alternative to the splines approach of Giorgi et al, as it provides estimates of the coefficients in time, analogous to the method od Therneau and Grambsch (2000). For testing goodness of fit, all our methods perform very well, which is especially evident through comparison to the splines approach.

Our work suggests that the proposed tests will be useful for samples of sizes 200 or more. Choice of which to use will depend upon what alternative model is considered more likely. If regression effects are expected to change between short and medium term survival then T_2 should be preferred. If the path of $\beta(t)$ is likely to be non-monotonic then T_3 should have greater power. This sort of situation can occur when there is some short-term instability in an acute phase immediately after diagnosis, followed by a chronic phase with

attenuating covariate effects. Overall T_1 seems to be a good choice. Of course further work on diagnostics for additive models is certainly necessary. For instance, one can foresee introduction of other versions of the tests (see also Kvaløy and Neef 2004), as the theory on random processes has plenty more to offer.

7 Software

All the procedures described in this paper were programmed in R by the second author, and are available upon request.

8 Appendix

Consider each subject's events to be an independent counting process $\{N_i(t), t \geq 0, i = 1, ..., n\}$ with intensity function (Andersen et al. 1993) given by

$$Y_i(t)(\lambda_P(t, X_{1i}) + \lambda_E^{(0)}(t)e^{\beta_0(t)X_i}),$$

where $Y_i(t)$ is the indicator process for i-th subject being at risk at time t. The true parameter is denoted by $\beta_0(t)$ to stress it does not have to be constant.

Define $U_i(\beta, t)$ and $\hat{E}(\beta, u)$ so that

$$\begin{aligned} U_i(\beta, t) := & \int_0^t \left\{ Z_i(u) - \hat{E}(\beta, u) \right\} dN_i(u) \\ = & \int_0^t \left(X_i - \sum_{j=1}^n X_j \frac{Y_j(u)(\lambda_P(u, X_{1j}) + \lambda_E^{(0)}(u)e^{\beta X_j})}{\sum_{j=1}^n Y_j(u)(\lambda_P(u, X_{1j}) + \lambda_E^{(0)}(u)e^{\beta X_j})} \right) dN_i(u) \end{aligned}$$

The partial residuals are defined for each person that has experienced an event:

$$U_i(\beta) := U_i(\beta, \infty) = X_i - \sum_{j \in R_i} X_j \frac{\left\{\lambda_P(t_i, X_{1j}) + \lambda_E^{(0)}(t_i) e^{\beta X_j}\right\}}{\sum\limits_{j \in R_i} \left\{\lambda_P(t_i, X_{1j}) + \lambda_E^{(0)}(t_i) e^{\beta X_j}\right\}}$$

The expression

$$\sum_{i=1}^{n} \left\{X_i(u) - \hat{E}(\beta, u)\right\} Y_i(u) (\lambda_P(u, X_{1i}) + \lambda_E^{(0)}(u) e^{\beta_0(u) X_i})$$

is equal to 0 at $\beta = \beta_0(u)$ for each time u and we can therefore express $U(\beta_0(t), t)$ as

$$U(\beta_0(t), t) = \sum_{i=1}^{n} \int_0^t \left\{X_i(u) - \hat{E}(\beta_0(u), u)\right\} dM_i(u),$$

where

$$M_i(t) = N_i(t) - \int_0^t Y_i(u) (\lambda_P(u, X_{1i}) + \lambda_E^{(0)}(u) e^{\beta_0(u) X_i}) du$$

is a martingale and $X_i(u) - \hat{E}(\beta_0(u), u)$ is a predictable process with respect to the filtration $\mathcal{F}_t = \sigma\{X_i, N_i(u), Y_i(u+) : 0 \leq u \leq t, i = 1, ...n\}$.

The residuals $U(\beta_0(t), t)$ can therefore be written as $\sum \int H_i dM_i$, where H_i is a predictable process and M_i is a martingale. It therefore holds (for each time t):

$$E\left(U(\beta_0(t), t)\right) = 0, \; E\left(U_i(\beta_0(t), t)\right) = 0 \tag{8}$$

and

$$\operatorname{cov}\left(U_i(\beta_0(t), t), U_j(\beta_0(t), t)\right) = E\left(\int_0^t H_i(u) dM_i(u), \int_0^t H_j(u) dM_j(u)\right) = 0.$$

The residuals thus have expected value 0 and are uncorrelated. Moreover, the variance of the residuals can be estimated by (Andersen et al. 1993)

$$V_i(\beta) := \frac{S^{(2)}(\beta, t_i)}{S^{(0)}(\beta, t_i)} - \left(\frac{S^{(1)}(\beta, t_i)}{S^{(0)}(\beta, t_i)}\right)^{\otimes 2},$$

where $S^{(r)}(\beta,t) = \sum_{i=1}^{n} X_i^{\otimes r} Y_i(t)(\lambda_P(t,X_{1i}) + \lambda_E^{(0)}(t)e^{\beta X_i})$ and for a column vector a, $a^{\otimes 2}$ denotes the outer product aa', $a^{\otimes 1}$denotes the scalar a and $a^{\otimes 0}$denotes the scalar 1.

References

Abrahamowicz, M., MacKenzie, T., Esdaile, J.M. (1996) Time-dependent hazard ratio: Modeling and hypothesis testing with application in lupus nephritis. *JASA,* **91**, 1432–1439

Andersen, P.K., Borch-Johnsen, K., Deckert, T., Green, A., Hougaard, P., Keiding, N. and Kreiner, S. (1985) A Cox regression model for relative mortality and its application to diabetes mellitus survival data. *Biometrics,* **41**, 921–932.

Andersen, P.K. and Væth, M. (1989) Simple parametric and nonparametric models for excess and relative mortality. *Biometrics,* **45**, 523–535.

Andersen, P.K, Borgan O., Gill, R.D., Keiding N. (1993) Statistical Models Based on Counting Processes. Springer, New York.

Bolard, P., Quantin, C., Esteve, J., Faivre, J., Abrahamowicz, M. (2001) Modelling time-dependent hazard ratios in relative survival: Application to colon cancer. *Journal of Clinical Epidemiology* **54**, 986–996.

Breslow, N.E., Lubin, J.H., Marek, P. and Langholz, B. (1983) Multiplicative models and cohort analysis. *Journal of the American Statistical Association*, **78**, 1–12.

Buckley, J.D. (1984) Additive and multiplicative models for relative survival rates. *Biometrics,* **40**, 51–62.

Csörgő, C., Faraway, J.J. (1996) The exact and asymptotic distributions of Cramér-von Mises statistics. *JRSS B*, **58**, 221–234.

Dickman, P.W., Sloggett, A., Hills, M., Hakulinen, T. (2004) Regression models for relative survival. *Statistics in Medicine,* **23**, 51–64.

Dickman, P.W., Hakulinen, T., Luostarinen, T., Pukkala, E., Sankila, R.,

Söderman, B., Teppo, L. Survival of cancer patients in Finland 1955–1994. *Acta Oncologica 1999*; **38**(Suppl. 12), 1–103.

Ederer, F., Axtell, L.M., Cutler, S.J. (1961) The relative survival rate: a statistical methodology. *National Cancer Institute Monograph,* **6**, 101-121.

Estève, J., Benhamou, E., Croasdale, M. and Raymond, M. (1990) Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine,* **9**, 529–538.

Giorgi, R., Abrahamowicz, M., Quantin, C., Bolard, P., Estève, J., Gouvernet, J., and Faivre, J. (2003) A relative survival regression model using B-spline functions to model non-proportional hazards. *Statistics in Medicine,* **22**, 2767–2784.

Hakulinen, T. and Tenkanen, L. (1987) Regression analysis of relative survival rates. *Applied Statistics,* **36**, 309-317. Kvaløy, J.T., Reiersølmoen Neef, L. (2004) Tests for the Proportional Intensity Assumption Based on the Score Process. *Lifetime Data Analysis*, **10**, 139–157.

O'Quigley, J. (2003) Khmaladze-type graphical evaluation of the proportional hazards assumption. *Biometrika,* **90**, 577-584.

O'Quigley, J., Stare, J. (2003) Cumulative empirical processes for survival models. In: Budin L, Luar-Stiffler V, Bekic Z, et al, editors. ITI 2003. Proceedings of the 25th international conference on Information technology interfaces; 2003 Jun 16-19; Cavtat. *Zagreb: SRCE, University of Zagreb*, 205–10.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.

Resnick, S.I. (1992) Adventures in Stochastic Processes. Birkhäuser, Boston.

Sasieni, P.D. (1996) Proportional excess hazards. *Biometrika*, **83**, 127–141.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239-241.

Stare, J., Henderson R., Pohar, M. (2005) An individual measure of relative survival. *JRSS C,* **54**, 115-126.

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