On an adaptive test of time-varying effects in Cox regression

Pierpaolo Brutti, Alessandra Nardi

LUISS Guido Carli and University of Rome "Tor Vergata"

Corresponding Author:
Pierpaolo Brutti, LUISS Guido Carli,
Dipartimento di Scienze Economiche e Aziendali, Viale Romania 32, 00197 Roma, Italy
E-mail: pbrutti@luiss.it

Summary
Cox's proportional hazards model is routinely used in many applied fields, especially in biomedical research. A common phenomenon in medical settings is the presence of a time dependency in the effect of one or more explanatory variables. In this situation, it is crucial to decide whether a covariate effect is constant, as prescribed by the standard Cox regression model, or not. Although the literature proposes several methods for estimating time-dependent effects in the Cox model, less attention has been paid to the problem of testing the null hypothesis of a proportional hazard against different possible alternatives. The main purpose of the present paper is to introduce a new test for time-varying effects in the proportional hazards model having power that adapts to the smoothness of the underlying function. Working on the Schoenfeld residuals our procedure is an adaptation to the present setting of a multiple testing technique introduced by Fromont and Laurent in 2006. The results are illustrated in reference to the well-known Mayo liver disease data.

KEY WORDS: *Cox model, goodness of fit, proportional hazard assumption, warped wavelets.*

Introduction

In survival analysis based on the well-known Cox's model (1), the hazard function for individual i is written as

$$\lambda_i\left(t\,|\,\beta, Z_i\right) = \lambda_0(t)\,\mathrm{e}^{Z_i^{\mathrm{T}}\beta}, \quad \forall\, i \in \{1,\ldots,n\}, \qquad [1]$$

where $\lambda_0\ (t)$ is a baseline hazard function, $\mathbf{Z}_i = \mathbf{Z}_i\ (t) = [Z_{i1}\ (t),\ldots,Z_{ip}\ (t)]^{\mathrm{T}}$ a vector of explanatory variables and $\boldsymbol{\beta} \in \mathbb{R}^p$ a vector of unknown parameters that describes the multiplicative effect of $\mathbf{Z}_i$ on $\lambda_0\ (t)$.

In its original formulation this model implies that the hazard ratio for any two individuals is independent of time. Yet, quite often, this assumption of proportional hazards is unrealistic and in contrast with the empirical evidence, especially in medical research. For example, the protective effect of a drug may diminish over time, or the increased hazard due to a risk factor may vanish after some years.

To inject flexibility into model [1], several non parametric procedures have been proposed for an extended Cox model with time-varying regression coefficients and possibly time-varying covariates, although in this paper we will focus on modelling time-dependent effects in one or more covariates, assuming $\mathbf{Z}_i$ to be constant over time

$$\lambda_i\left(t\,|\,\beta(t), \mathbf{Z}_i(t)\right) = \lambda_0(t)\,\mathrm{e}^{\mathbf{Z}_i(t)^{\mathrm{T}}\beta(t)}. \qquad [2]$$

Some early examples are the methods based on regression estimators proposed by Murphy and Senn (2), Hess (3) and Abrahamowicz et al. (4). Different estimators for $\boldsymbol{\beta}(t)$ have also been developed using penalized partial likelihood (5-7) and local partial likelihood (8-10) approaches. Alternatively, residuals can be used to detect and model non proportionality, as in (11-14).

It can be noted that the vast literature available focuses mainly on estimating time-varying effects, while less attention has been devoted to testing them [see (15) for a data-driven smooth test related to our proposal; section 6.2 of (16) and (17) for nice reviews]. However, point estimates, no matter how good, are of little use without some indication of the magnitude of random variation.

Our aim here is to propose a flexible non parametric test for proportional hazards having power that adapts to the smoothness of the underlying function $\boldsymbol{\beta}(t)$. To this end, we follow a residuals-based approach in which Schoenfeld residuals play a central role. Our procedure is an adaptation to the present setting of a multiple testing technique introduced by Fromont and Laurent in (18).

The outline of the paper is as follows: after an introduction to the Schoenfeld residuals and their use in checking and understanding proportional hazards, we describe our testing procedure. An application to the well-known Mayo liver disease data is then described, before our final remarks. All the relevant technical details are collected in the Appendix.

Schoenfeld residuals and proportional hazards testing

Schoenfeld in (19) introduced a class of residuals for Cox's model directly linked to the partial likelihood. They are defined as follows: let $\Re_k$ be the *risk set* corresponding to the observed failure time t_k with $k \in \{1,...,\mathrm{d}\}$, and $\hat{\boldsymbol{\beta}}$ a vector of estimated parameters. Thus, when there are no tied event times, the Schoenfeld residuals $\{\mathbf{r}_k\}_k$ are given by

$$\mathbf{r}(t_k) = \mathbf{r}_k = \mathbf{Z}_{(k)} - \overline{\mathbf{z}}\left(\hat{\beta}, t_k\right) \quad \text{with} \\ \overline{\mathbf{z}}(\beta, t_k) = \sum_{i \in \Re_k} \overline{w}_i(\beta, t_k) \cdot \mathbf{Z}_i(t_k), \qquad [3]$$

where $\mathbf{Z}_{(k)}$ denotes the covariate vector for the individual observed to fail in t_k, and $\overline{\mathbf{z}}\,(\hat{\boldsymbol{\beta}},t_k)$ is just the mean of $\mathbf{Z}$ over those subjects still at risk at t_k weighted by their normalized risk scores

$$\overline{w}_i\left(\hat{\beta}, t_k\right) = \frac{\mathrm{e}^{\mathbf{Z}_i(t_k)^{\mathrm{T}}\hat{\beta}(t_k)}}{\sum_{j \in \Re_k} \mathrm{e}^{\mathbf{Z}_j(t_k)^{\mathrm{T}}\hat{\beta}(t_k)}}, \quad \forall i \in \Re_k. \qquad [4]$$

A Schoenfeld residual can be regarded as a *distance* between the covariate vector of an individual observed to fail in t_k and the expected covariate vector, i.e. the vector characterizing that patient, among those still at risk in t_k, who is expected to fail according to the Cox model. From this perspective, Schoenfeld residuals resemble the classic definition of residuals.

Some features of Schoenfeld residuals make them especially suitable for diagnosing the presence of time-dependent effects. For every deceased individual a vector of residuals is defined, one for each explanatory variable; this makes it possible to check each covariate separately and specifically for a possible time-varying effect.

When proportionality holds true, Schoenfeld residuals have no systematic pattern over time and a smoothed plot of the ℓ–th component $\mathrm{r}_{\ell,k}$ of $\mathbf{r}_k$ against t_k is expected to be centred about 0. This original proposal was subsequently refined by several authors.

Grambsch and Therneau (11) noted that spurious time-dependent effects could be detected due to correlated covariates and proposed scaling Schoenfeld residuals by their weighted covariance matrix (see Appendix A for details).

Taking $\mathbf{r}_k^*$ as the scaled Schoenfeld residuals, the same authors (11) showed that

$$\mathrm{E}(\mathbf{r}_k^*) + \hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}(t_k). \qquad [5]$$

This suggests that a smoothed plot of the ℓ–th component $\mathrm{r}_{k,\ell}^* = \mathrm{r}_\ell^*(t_k)$ of $\mathbf{r}_k^*$ against t_k can be used not only to diagnose the presence of non proportionality for covariate $\ell \in \{1,...,p\}$, but also to reveal the form of a possible time-varying effect.

In other words, residuals can be scaled, smoothed and added to the initial constant estimate to obtain a crude estimate of $\boldsymbol{\beta}(t)$ in model [2]. This approach has advantages over direct non parametric estimation as, on the one hand, estimates are easier to calculate, to interpret and to relate to the more restricted Cox model estimate, and on the other, it is also the basis for improved, consistent estimators like the one based on iterated residuals proposed by Winnet and Sasieni in (14).

Thus, equation [5] is a powerful tool for exploring and validating the proportionality assumption. Typically a line can be fitted to the plot and this can be followed by a test for a zero slope; a non zero slope is evidence against proportional hazards. Interestingly, most of the tests for proportional hazards that have been proposed in the literature are of the same type: Rao efficient score tests of $\mathrm{H}_0 : \theta = 0$ under the model

$$\beta_\ell(t) = \beta_\ell + \theta_\ell(\mathrm{g}_\ell(t) - \overline{\mathrm{g}}_\ell), \text{ with} \\ \overline{\mathrm{g}}_\ell = \text{average } (\{g_\ell(t_k)\}_k), \text{ and } \ell \in \{1,...,p\} \qquad [6]$$

for different choices of the time transform $\mathbf{g}(t)$.

The ability to "see" the test is clearly an advantage but it comes with a few drawbacks. First of all, it can be noticed that the weighted variance matrix used to scale Schoenfeld residuals, being a weighted variance of the covariates for individuals still at risk at t_k, may become unstable as individuals are removed from the risk set $\Re_k$ because of death or censoring. A possible solution is to control this instability in the estimate for late failure times (when the risk set is substantially reduced) using a weighted smoother, as in (13).

A more serious problem, however, is that there are forms of non proportionality that the family of tests discussed above may completely fail to detect.

A simple example is a quadratic shape for $\boldsymbol{\beta}(t)$: it might be apparent on the plot, but be totally missed by the test of linear slope. To solve this issue we might consider more general tests. For example, we might fit a quadratic function to the plot and then conduct a two degree of freedom test; or maybe consider a higher degree fit with the associated test, but the point is that the need to pick in advance an appropriate time transform $\mathbf{g}(\cdot)$ and a reasonable *parametric* model under the null is, ultimately, unsatisfactory. Such a specification will rarely be known a priori and the common practice of basing the choice of a parametric function on residuals from the Cox model and of testing it using the same data set is not correct, resulting in a double use of sampling information. For this reason, we here suggest recasting the problem in a non parametric regression framework, modelling the scaled residuals as an unknown smooth function $f(\cdot)$ plus (bounded) noise:

$$\mathrm{r}^*_\ell(t) = f(t) + \varepsilon(t),$$

and then applying a suitable goodness-of-fit procedure to test $\mathrm{H}_0 : f \equiv f_0$ against $\mathrm{H}_1 : f \neq f_0$. Given the nature of the data, it seems reasonable to look for a test that has power only against smooth alternatives and that ideally is adaptive to the (unknown) smoothness of the underlying function [see (20)]. The next section describes such a test.

A goodness-of-fit test

As mentioned in the previous section, the framework we shall work with in this paper is the usual non parametric regression problem with random design. In this model we observe an i.i.d. sample $D_{1:d} = \{\mathbf{D}_k = (T_k, Y_k)\}_{k \in \{1...d\}}$ from the distribution of a vector $\mathbf{D} = (T, Y)$ described structurally as

$$Y = f(T) + \varepsilon,$$

where T and ε denote respectively the design variable and the stochastic error term. Notice that in our case $Y_k = \mathrm{r}^*_\ell(t_k)$ for some $\ell \in \{1,...,p\}$ of interest. Our aim is to test $\mathrm{H}_0 : f \equiv f_0$ against $\mathrm{H}_1 : f \neq \mathrm{f}_0$ where usually $f_0 = 0$. In this section we will provide a qualitative description of our method, while leaving the details to Appendix B.

In general, to tackle a testing problem effectively we need to:

1. choose a suitable *distance*[1], say $d(\cdot,\cdot)$, to measure departures of the parameter of interest $f(\cdot)$ from the null "guess" $f_0(\cdot)$;
2. define a test statistic based on a nicely behaved estimator $\hat{d}$ of d;
3. find a way of quantifying the random fluctuations of $\hat{d}$ in order to control the error rates of the test.

In our setting the parameter of interest is the unknown function $f(\cdot)$, thus one natural possibility is to take

$$\tilde{d}\left(f, \mathrm{H}_0\right) = \int \left| f(t) - f_0(t) \right|^2 \mathrm{d}t,$$

but in this way we completely neglect the distortion induced by the design that we assumed to be random or, at least, non equispaced being induced by the observed failure times $\{t_k\}_k$. To fix this, let $G_T(\cdot)$ be the distribution function of the design variable T, then define

$$d\left(f, \mathrm{H}_0\right) = \int \left| f(t) - f_0(t) \right|^2 G_T(\mathrm{d}t) = \left\| f - f_0 \right\|^2_{G_T}. \quad [7]$$

Roughly speaking, d is the weighted version of $\tilde{d}$ with the weights given by the probability of observing any particular design value t.

Now that we know how to measure deviations from H_0 in the population, we are left with the problem of estimating this metric from the data. An intuitive option would be to estimate f first with an estimator $\hat{f}$, and then plug it in equation 7 obtaining as the test statistic $\hat{d} = d\,(\hat{f}, \mathrm{H}_0)$. Although feasible, this path is, in a sense, rather cumbersome. To understand why, let f_0 be equal to zero on the whole domain. In this

[1] The function d does not need to be a distance in a technical sense. It is enough that, in the population, it takes on a unique and distinctive value only when $f \equiv f_0$.

way, according to step 2 above, the quantity we should really estimate is $d\,(f,\mathrm{H}_0) = \|f\|^2_{GT}$, a real non negative number[2], not a function! In addition we would like to find an estimator $\hat{d}$ capable of recovering the complexity of the underlying function from the data and exploiting this information to improve on the power of the test whenever possible.

To achieve all this we follow (18) and disassemble the original level–α test in a sequence of "subtests" of growing complexity for $\mathrm{H}_0 : f = f_0$, rejecting the "grand-null" if, for some of the tests in the collection, the hypothesis is rejected. Notice that all the tests in the collection share the same null hypothesis; it is the test statistic that varies in a controlled fashion.

More precisely, taking $f_0 = 0$ once again, for each $m \in \{0,...,\mathrm{M}\}$, the m–th subtest is characterized by the following steps:

- define a "probing set" S_m for $f(\cdot)$ such that $S_{m-1} \subset S_m$ and $S_0 = \{0\}$;
- consider the *projection* $\Pi_{S_m}(f)$ of $f(\cdot)$ onto S_m;
- reject $\mathrm{H}_0 : f = 0$ if an estimator $\hat{d}_m$ of $d_m\,(f,\mathrm{H}_0) = \|\Pi_{S_m}(f)\|^2_{GT}$ is greater than an appropriate quantile[3].

The idea behind this procedure is relatively easy: each probing set S_m captures some features of the function f and by projecting we extract more and more complex structures from f as m increases. Clearly the choice of M is crucial and strictly linked to how we build the *sieve* $\{S_m\}_m$. In this regard, the reader is invited to refer to the Appendix for further comments. Here, instead, we have to settle one last question: what kind of estimator $\hat{d}_m$ should we use at each step, and consequently how can we quantify its sampling distribution under the null? As we shall see in Appendix B, $\hat{d}_m$ belongs to the class of *U-statistics* and, rather luckily, some well-known results can be applied to evaluate its variability [see (21)].

Application to Mayo liver disease data

To illustrate our proposal we consider well-known primary biliary cirrhosis (PBC) data [see (22)] which refer to a total of 418 patients followed until death or censoring. The data come from a Mayo Clinic trial of PBC of the liver, conducted between 1974 and 1984. PBC is a progressive disease thought to be of an autoimmune origin; it is associated with an inflammatory process that eventually leads to cirrhosis and the death of the patient. We investigate the predictive effect on survival of the following covariates: albumin (mg\Dl), bilirubin (mg\Dl), oedema (present\non present), prothrombin time (seconds) and age (years). In order to obtain a more stable fit, the logarithmic transformation was applied to bilirubin and prothrombin time, both the covariates showing an extremely skew distribution Results from a standard Cox regression model are reported in Table 1.

All effects were clearly significantly different from 0. After fitting the Cox model, an appropriate investigation for departures from proportionality is recommended. A smoothed plot of Schoenfeld residuals against the observed failure times provides an initial graphical representation. However any graphical procedure, to be of real help, should be accompanied by a measure of the empirical evidence against the null hypothesis. We compare the adaptive test we propose with two methods that are commonly used in medical research. The first is a test based on scaled Schoenfeld residuals that was introduced in 1994 by Grambsch and Therneau (11) (GT). This method has the advantage of being in agreement with the graphical evaluation; however, belonging to the class of Rao efficient score tests we discussed in section 2, it may fail to detect non linear forms of non proportionality.

Table 1. Results for the fit of a five-parameter Cox model to the PBC data.

	β	exp(β)	S.E.	z	p
Age	0.0382	1.039	0.00767	4.98	<0.001
Albumin	-0.7385	0.478	0.21029	-3.51	<0.001
log(Bilirubin)	0.8975	2.454	0.08279	10.84	<0.001
Oedema	0.6661	1.947	0.20635	3.23	0.001
log(Prothrombin time)	2.3314	10.293	0.77360	3.01	0.003

[2] The function d is actually a quadratic *functional* of f.

[3] The level of the overall test has to be equal to a given $\alpha \in (0,1)$.

A different approach is based on the introduction into the Cox model of interactions between each covariate and a suitable function of time. We remark that the specification of such a function, rarely known in advance, cannot be based on the data set to be analyzed. In order to avoid the choice of a specific functional form, Hess (3), in 1994, proposed the use of restricted cubic splines (RCS). Implemented as time-by-covariate interactions, standard methods and statistical software can be used for testing regression coefficients. In particular, a test for detecting a time-dependent effect can be based on the pertinent set of parameters, adjusting for the appropriate degrees of freedom. The disadvantage of this method is that the researcher must choose the number and the position of the knots, this choice being somewhat arbitrary but, at the same time, influential on the final result. We assumed three and five knots, located at the percentiles of the failure time distribution; 25%, 50% and 75% in the former case, 10%, 25%, 50%, 75% and 90% in the latter. It must be noted that, as the assumptions of a constant effect concern every single covariate, separate tests should be done. In our case this is not a serious issue; however, in the presence of several explanatory variables, a correction for multiplicity might be necessary.

The results of the different approaches are summarized in Table 2.

Only prothrombin time and the presence of oedema showed a systematic departure from the assumption of a constant effect.

Figure 1 shows a plot of scaled Schoenfeld residuals for prothrombin time versus observed failure times. A spline-based non parametric estimate suggests that prothrombin time leads to an initially increased risk which wears off towards the end of the follow up. Application of the proposed test in order to verify whether this observed behaviour can still be consistent with a constant hazard ratio, a *p*–value <0.001 confirms a significant departure from proportionality. A similar message comes from the other tests but with considerably lower strength.

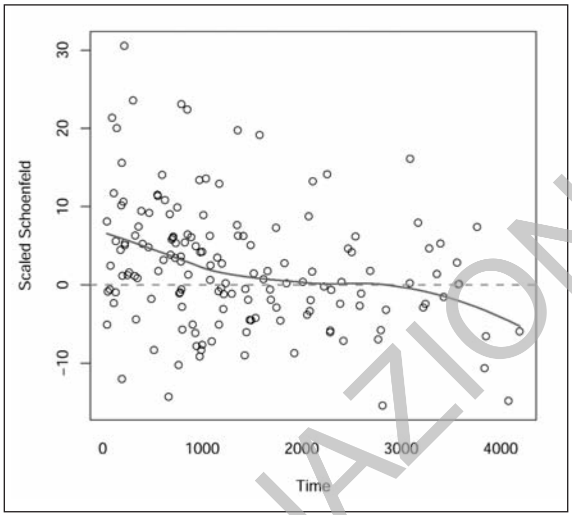

Figure 1. PBC data, time-dependent coefficient plot for prothrombin time.

A similar decaying effect is shown in Figure 2 for the

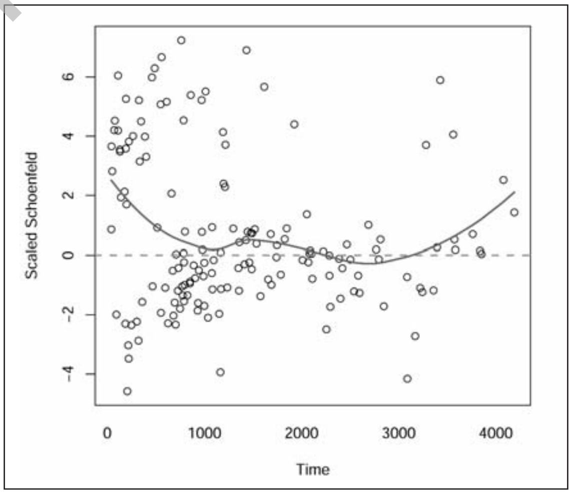

Figure 2. PBC data, time-dependent coefficient plot for oedema.

Table 2. *p*-values resulting from the different approaches.

	GT	RCS (3 knots)	RCS (5 knots)	Adaptive
Age	0.953	0.915	0.639	0.713
Albumin	0.783	0.942	0.772	0.901
log(Bilirubin)	0.269	0.286	0.002	0.117
Oedema	0.042	0.011	0.048	0.025
log(Prothrombin time)	0.003	0.005	0.022	< 0.001

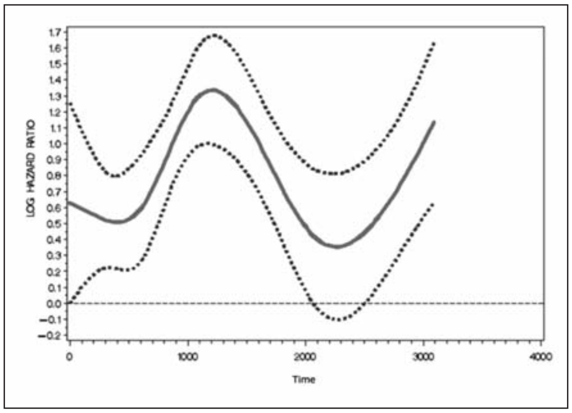

Figure 3. PBC data, time–dependent coefficient plot for bilirubin based on five-knot restricted cubic splines.

presence of oedema even though the empirical evidence appears to be less strong than in the previous case. Again, the hypothesis of a time-varying effect is supported by a *p*–value of 0.02, higher than 0.001 but still significant.

The *p*–values from the four different methods in Table 2 are of different magnitude but show a consistent pattern. The only exception is a significant value for bilirubin when RCS are used, assuming five knots.

Note that the two tests based on RCS require a preliminary estimate of the time-varying effects. Especially when five knots are assumed, this approach may result in a considerable amount of over-fit due to the inclusion of unnecessary parameters for the possible time dependence of each single covariate. Figure 3 shows the spurious time-varying effect for the logarithmic transformation of bilirubin. A *p*–value of 0.0021 suggests a significant departure from proportionality. However, none of the other tests confirms these results and the estimated time-dependent effect is rather unrealistic from a clinical perspective.

Finally, with respect to the GT test our proposal has the advantage of being sensitive to non linear departures from proportional hazards, at the same time guaranteeing a power that adapts to the smoothness of the underlying time-dependent effect.

Discussion

Testing for proportional hazards is a crucial problem in many applied settings. In spite of this, the vast majority of the tests available are quite restrictive and parametric in nature. The present paper tackles this problem from a residuals-based point of view, recasting it in a non parametric regression framework in order to add flexibility to the usual techniques. The result is an effective non parametric test of time-varying effects that do not require new or demanding computational tools and that is based on a procedure having power that adapts to the smoothness of the underlying function.

The preliminary numerical study based on the PBC data contained in section 4 suggests that our method is on the right track but an extensive simulation study coupled with a deeper theoretical understanding of its performance in the present setting is strongly needed. We will elaborate more on this in future research, as well as considering other classes of residuals [e.g. the iterated Schoenfeld introduced in (14)] and of basis functions. We also hope to discuss the power of the test against very smooth/convex alternatives, the most natural ones in a biomedical context.

References

1. Cox DR. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B 1972; 34: 187-220.
2. Murphy SA, Senn PK. Time-dependent coefficients in a Cox-type regression model. Stochastic Processes and Applications 1991; 39:153-180.
3. Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. Stat Med 1994; 13:1045-1062.
4. Abrahamowicz M, MacKenzie T, Esdaile JM. Time-dependent hazard ratio: modelling and hypothesis testing with application in lupus nephritis. J Am Stat Assoc 1996, 91:1432-1439.
5. Hastie T, Tibshirani R. Varying-coefficient models (with discussion). Journal of the Royal Statistical Society, Series B 1993, 55:757-796.
6. Verweij JM, van Houwelingen HC. Time-dependent effects of fixed covariates in Cox regression. Biometrics 1995;51:1550-1556.
7. Zucker DM, Karr AF. Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. Annals of Statistics 1990; 18:329-353.
8. Bown D, Kauermann G, Ford I. A partial likelihood approach to smooth estimation of dynamic covariate effects using penalised splines. Biom J 2007; 49: 441-452.

9. Cai Z, Sun Y. Local linear estimation for time-dependent coefficients in Cox's regression models. Scand J Stat 2003, 30: 93-111.
10. Valsecchi MG, Silvestri D, Sasieni P. Evaluation of long-term survival: use of diagnostics and robust estimators with Cox's proportional hazards model. Stat Med 1996; 15:2763-2780.
11. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994; 81: 515-526. Correction, 82: 668.
12. Scheike TH, Martinussen T. On estimation and tests of time-varying effects in the proportional hazards model. Scand J Stat 2004; 31: 51-62.
13. Winnett A, Sasieni P. A note on scaled Schoenfeld residuals for the proportional hazards model. Biometrika 2001; 88:565-571.
14. Winnett A, Sasieni P. Iterated residuals and time-varying covariate effects in Cox regression. Journal of the Royal Statistical Society, Series B. 2003; 65: 473-488.
15. Kraus D. Data-driven smooth tests of the proportional hazards assumption. Lifetime Data Anal 2007; 13: 1-16.
16. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health. Springer New York, 2000.
17. Kvaløy JT, Neef LR. Tests for the proportional intensity assumption based on the score process. Lifetime Data Anal 2004; 10: 139-157.
18. Fromont M, Laurent B. Adaptive goodness-of-fit tests in a density model. Annals of Statistics 2006; 34: 680-720.
19. Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika 1982; 69: 239-241.
20. Ingster Y, Suslina IA. Nonparametric Goodness-of-Fit Testing Under Gaussian Models, first edition. Springer-Verlag New York, 2002.
21. de la Peña VH, Giné E. Decoupling: From Dependence to Independence. Probability and its Applications. Springer-Verlag New York, 1999.
22. Fleming TR, Harrington DP. Counting Processes and Survival Analysis. Wiley-Interscience Hoboken, New Jersey, 1991.
23. Cucker F, Smale S. On the mathematical foundations of learning. Bulletin of the American Mathematical Society 2002; 39:1-49.
24. Hart JD. Nonparametric Smoothing and Lack-of-Fit Tests. Springer Series in Statistics. Springer-Verlag New York, 1997.
25. Laurent B. Adaptive estimation of a quadratic functional of a density by model selection. ESAIM: Probability and Statistics 2005; 9:1-18.
26. Shao J, Tu D. The Jacknife and Bootstrap. Springer-Verlag New York, 1995.
27. Brutti P. New confidence regions for nonparametric regression and how to explore them. PhD thesis, Department of Statistics, Probability and Applied Statistics, University of Rome "La Sapienza", 2005.
28. Kerkyacharian G, Picard D. Thresholding in learning theory. Constructive Approximation 2007; 26: 173-203.
29. Kerkyacharian G, Picard D. Regression in random design and warped wavelets. Bernoulli 2004, 10:1053-1105.
30. Mallat S. A Wavelet Tour of Signal Processing, second edition. Academic Press London, 1998.

Appendix

A. Scaled Schoenfeld residuals

Grambsch and Therneau (11) proposed scaling Schoenfeld residuals in order to avoid the detection of spurious time-dependent effects. Let

$$\Sigma(\boldsymbol{\beta},t_k) = \sum \bar{w}_i\,(\boldsymbol{\beta},t_k) \cdot \|\mathbf{Z}_i\,(t_k) - \bar{\mathbf{z}}\,(\boldsymbol{\beta},t_k)\|_2^2, \qquad [8]$$

be the weighted variance matrix of $\mathbf{Z}$ at time t_k. If $\hat{\boldsymbol{\beta}}$ denotes the coefficient from an ordinary fit of the Cox model, then we can define the *scaled Schoenfeld residuals* $\{\mathbf{r}_k^*\}_k$ as follows

$$\mathbf{r}_k^* = \Sigma^{-1}\,(\hat{\boldsymbol{\beta}},t_k)\cdot \mathrm{r}_k. \qquad [9]$$

B. A goodness-of-fit test: technical details

We observe an i.i.d. sample $D_{1:d} = \{\mathbf{D}_k = (T_k,Y_k)\}_{k\in\{1\ldots\mathrm{d}\}}$ from the distribution of a vector $\mathbf{D} = (T,Y)$ described structurally as

$$Y = f\,(T) + \varepsilon,$$

for (T,ε) a random vector with $\mathrm{E}(\varepsilon|T) = 0$ and $\mathrm{E}(\varepsilon|^2T) < \infty$ (almost certainly). Notice that in our case $Y_k = \mathrm{r}_\ell^*\,(t_k)$ for some $\ell \in \{1,\ldots,p\}$ of interest. The regression function is known to belong to a subset F of $\mathrm{L}^2\,([0,1],G_T)$, G_T being the marginal distribution of T. We do not assume that the errors are normally distributed, and we do not assume that T and ε are independent but, mainly for technical reasons, we will assume, as in the majority of the current literature on learning theory [see (23)], that $|f\,(t) - y|$ is uniformly bounded (almost everywhere) by a positive constant C.

As is often the case in non parametric statistics, we could cast this example into a problem of estimating a sequence $\theta = [\theta_1,\theta_2,\ldots] \in \ell^2$ of parameters by expanding $f(\cdot)$ on a fixed orthonormal basis $\{e_j\}_{j\in\mathbb{N}}$ of $\mathrm{L}^2\,([0,1],G_T)$. The Fourier coefficients take the form

$$\theta_j = \left\langle f,e_j \right\rangle_{\mathrm{L}^2(G_T)} = \mathrm{E}_{(T,Y)}\left[Y \cdot e_j(T)\right],$$

and they can be estimated unbiasedly by

$$W_j = \tfrac{1}{d}\sum_{k=1}^{d} Y_k \, e_j(T_k),$$

even though it does not appear particularly useful to move directly in sequence space by considering $[W_1, W_2, ...]$ as the observation vector. What we propose is a goodness-of-fit test similar to the one introduced in (18) [see (24) for a nice review of nonparametric lack-of-fit tests]. To describe it, let $f_0(\cdot)$ be some fixed function in $\mathrm{L}^2([0,1], G_T)$ and $\alpha \in (0,1)$. Now let us suppose that our goal is to build a level–α test of the null hypothesis $\mathrm{H}_0 : f \equiv f_0$ against the alternative $\mathrm{H}_1 : f \neq f_0$ from the data $\{\mathbf{D}_i\}_{i\in\{1,...,d\}}$. The test is based on the estimation of

$$\left\| f - f_0 \right\|^2_{\mathrm{L}^2(G_T)} = \left\| f \right\|^2_{\mathrm{L}^2(G_T)} + \left\| f_0 \right\|^2_{\mathrm{L}^2(G_T)} - 2\left\langle f, f_0 \right\rangle_{\mathrm{L}^2(G_T)}.$$

Since the last (linear) term $\left\langle f, f_0 \right\rangle_{\mathrm{L}^2(G_T)}$ can be estimated easily by the empirical estimator $\tfrac{1}{n}\sum_{k=1}^{d} Y_k \, f_0(T_k)$, the key problem is the estimation of the first term $\|f\|^2_{\mathrm{L}^2(G_T)}$. Adapting the arguments in (25), we can consider an at most countable collection of linear subspaces of $\mathrm{L}^2([0,1], G_T)$ denoted by $\mathrm{S} = \{S_m\}_{m\in\{1,...,\mathrm{M}\}}$. For all $m \in \{1,...,\mathrm{M}\}$, let $\{e_j\}_{j\in I_m}$ be some orthonormal basis of S_m. The estimator

$$\hat{\theta}_{d,m} = \frac{1}{d(d-1)} \sum_{k_1=2}^{d} \sum_{k_2=1}^{d-1} \left[\sum_{j\in I_m} \left\{ Y_{k_1} e_j(T_{k_1}) \right\} \cdot \left\{ Y_{k_2} e_j(T_{k_2}) \right\} \right] =$$

$$= \frac{1}{d(d-1)} \sum_{k_1=2}^{d} \sum_{k_2=1}^{d-1} h_m(\mathbf{D}_{k_1}, \mathbf{D}_{k_2}),$$

is a U-statistic of order two [see (21)] for $\left\| \Pi_{S_m}(f) \right\|^2_{\mathrm{L}^2(G_T)}$ – where $\prod_{S_m}(\cdot)$ denotes the orthogonal projection onto S_m – with kernel

$$h_m(\mathbf{d}_1, \mathbf{d}_2) = \sum_{j\in I_m} \left\{ y_1 e_j(t_1) \right\} \cdot \left\{ y_2 e_j(t_2) \right\},$$

$$\mathbf{d}_k = (t_k, y_k), \; k \in \{1, 2\}.$$

Thus, for any $m \in \{1,...,\mathrm{M}\}$, $\|f - f_0\|^2_{\mathrm{L}^2(G_T)}$ can be estimated by

$$\hat{\mathrm{L}}_{d,m} = \hat{\theta}_{d,m} + \left\| f_0 \right\|^2_{\mathrm{L}^2(G_T)} - \frac{2}{n} \sum_{i=1}^{n} Y_i \, f_0(T_i). \qquad [11]$$

Now that we have an estimator $\hat{\mathrm{L}}_{d,m}$, let us denote by $l_{d,m}(u)$ its $1 - u$ quantile under H_0, and consider

$$u_\alpha = \sup\left\{ u \in (0,1) : \mathrm{P}^{\otimes d}_{f_0} \left[\sup_m \left\{ \hat{\mathrm{L}}_{d,m} - l_{d,m}(u) \right\} > 0 \right] \leq \alpha \right\},$$

where $\mathrm{P}^{\otimes d}_{f0}\{\cdot\}$is the law of the observations $\{\mathbf{D}\}_{i\in\{1,...,d\}}$ under the null hypothesis. Then introduce the test statistic L_α defined by

$$\mathrm{L}_\alpha = \sup_m \left\{ \hat{\mathrm{L}}_{d,m} - l_{d,m}(u_\alpha) \right\},$$

so that we reject the null whenever L_α is positive.

This method, adapted to the regression setting using [9], amounts to a multiple testing procedure. Indeed, for all $m \in \{1,...,\mathrm{M}\}$, we construct a level–u_α test by rejecting $\mathrm{H}_0 : f \equiv f_0$ if $\hat{\mathrm{L}}_{d,m}$ is greater than its $(1 - u_\alpha)$ quantile under H_0. After this, we are left with a collection of tests and we decide to reject H_0 if, for some of the tests in the collection, the hypothesis is rejected. In practice, the value of u_α and the quantile $\{l_{d,m}(u_\alpha)\}_m$ are to be estimated by a *wild bootstrap* procedure [see (26)] as explained in (27).

Warped wavelets

Both the practical and theoretical performances of the proposed test depend strongly on the orthogonal system we adopt to generate the collection of linear subspaces $\{S_m\}_m$. In dealing with a density model, Fromont and Laurent (18), consider a collection obtained by mixing spaces generated by constant piecewise functions (Haar basis), scaling functions from a wavelet basis, and, in the case of compactly supported densities, a trigonometric polynomial. Clearly these bases are not orthonormal in our weighted space $\mathrm{L}^2([0,1], G_T)$, hence we have to consider other options.

A basis that proved to fit perfectly in the present framework is the so-called *warped wavelet* basis studied by Kerkyacharian and Picard in (28, 29). The idea is as follows. For a signal observed at some design points, $Y(\mathrm{t}_k)$, $k \in \{1,...,2^j\}$, if the design is regular ($\mathrm{t}_k = k/2^J$), the standard wavelet decomposition algorithm starts with $s_{J,k} = 2^{J/2}\, Y(k/2^J)$ which approximates the scaling coefficient $\int Y(t)\, \phi_{J,k}(t)\, \mathrm{d}t$, with $\phi_{J,k}(t) = 2^{J/2}\, \phi(2^J - k)$ and $\phi(\cdot)$ the so-called scaling function or father wavelet. Then the cascade algorithm is employed to obtain the wavelet coefficients $d_{j,k}$ for $j \leq J$, which in turn are thresholded [see (30) for further information]. If the design is not regular, and we still employ the *same* algorithm, then for a function $H(\cdot)$ such that $H(k/2^J) = t_k$, we have $s_{J,k} = 2^{J/2}\, Y\left(\mathrm{H}(k/2^J)\right)$. Essentially what we are doing is decomposing, with respect to a standard wavelet basis, the function $Y\left(H(\mathrm{t})\right)$ or, if $G \circ H(t) \equiv t$, the original function $Y(x)$ itself but with respect to a new *warped* basis $\left\{ \psi_{j,k}\left(G(\cdot)\right) \right\}_{(j,k)}$. In the regression setting, this means replacing the standard wavelet expansion of the function $f(\cdot)$ with its expansion on the new basis $\left\{ \psi_{j,k}\left(G(\cdot)\right) \right\}_{(j,k)}$, where $G(\cdot)$ is adapting to the design: it may be the distribution function of the design $G_T(\cdot)$, or its estimation $\hat{G}_T(\cdot)$ when the distribution function is unknown. An appealing feature of this method is that it does not need a new algorithm to be implemented: just standard and widespread tools.

The rate of optimality and adaptivity of the procedure that results from coupling warped wavelets and the test described above can be found in (27) and depends quite heavily on how we choose J or, using the notation of section 3, M. There are different ways of assessing this issue [see (18) and (24)], but in practice we have seen that a value of M between 10 and 20 (depending on the sample size) works reasonably well.