



Focused Information Criteria and Model Averaging for the Cox Hazard Regression Model

Nils Lid Hjort & Gerda Claeskens

To cite this article: Nils Lid Hjort & Gerda Claeskens (2006) Focused Information Criteria and Model Averaging for the Cox Hazard Regression Model, Journal of the American Statistical Association, 101:476, 1449-1464, DOI: [10.1198/016214506000000069](https://doi.org/10.1198/016214506000000069)

To link to this article: <https://doi.org/10.1198/016214506000000069>



Published online: 01 Jan 2012.



Submit your article to this journal [↗](#)



Article views: 564



View related articles [↗](#)



Citing articles: 15 View citing articles [↗](#)

Focused Information Criteria and Model Averaging for the Cox Hazard Regression Model

Nils Lid HJORT and Gerda CLAESKENS

This article is concerned with variable selection methods for the Cox proportional hazards regression model. Including excessive covariates causes extra variability and inflated confidence intervals for regression parameters; thus regimes for discarding the less informative ones are needed. Our framework has p covariates designated as “protected,” while variables from a further set of q covariates are examined for possible inclusion or exclusion. We develop a focused information criterion (FIC) that for given interest parameter finds the best subset of covariates. Thus the FIC might find that the best model for predicting median survival time is different than the best model for estimating survival probabilities, and the best overall model for analyzing men’s survival might not be the same as the best overall model for analyzing women’s survival. Methodology is also developed for model averaging, wherein the final estimate of a quantity is a weighted average of estimates computed for a range of submodels. Our methods are illustrated in simulations and for a survival study of Danish skin cancer patients.

KEY WORDS: Akaike information criterion; Covariate selection; Cox regression; Focused information criteria; Median survival time; Model averaging.

1. INTRODUCTION

Suppose that survival data of the form $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$ are recorded for n individuals, where t_i is lifetime (possibly censored); δ_i is an indicator for noncensoring; $\mathbf{x}_i$ contains, say, p covariates that are deemed necessary in the regression model; and $\mathbf{z}_i$ has, say, q further covariates potentially worthy of inclusion. The most popular model for such data is the Cox model of proportional hazards, where the hazard rate for individual i is expressed as

$$h_i(u) = h_0(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\gamma}) \quad \text{for } i = 1, \dots, n. \quad (1)$$

Here $h_0(u)$ is assumed to be continuous and positive over the range of lifetimes of interest but is otherwise not specified. This makes the model partly parametric and partly nonparametric. Inference about $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ typically proceeds using the well-known partial likelihood $L_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$, which is properly defined in Section 2.

This article is concerned with developing methods for selecting covariates $z_{i,j}$ that are, in some sense, best among the q . The argument against simply including all of them is that this may cause excessive estimation variability, leading to inflated confidence intervals and less-powerful tests. On the other hand, including too few covariates could mean serious modeling bias and missing important explanatory features in the analysis. Thus selecting the “best” set is a statistical balancing act between bias and variance.

1.1 The Danish Malignant Melanoma Study

For an illustration (revisited in Sec. 9), we study the skin cancer survival analysis dataset described and extensively analyzed by Andersen, Borgan, Gill, and Keiding (1993) and others. In this Danish study, 205 patients with malignant melanoma underwent radical removal surgery and were followed after

surgery over the period of 1962–1977. Several covariate variables are of potential interest for studying survival chances, including x_1 , an indicator for sex of the patient; z_1 , thickness of the tumor; z_2 , infection infiltration level, a measure of resistance against the tumor; z_3 , presence so-called “epithelioid cells”; z_4 , presence of ulceration; z_5 , invasion depth (at levels 1, 2, and 3); and z_6 , age of the patient at the operation (in years). Among the findings of Andersen et al. (1993) were that men tend to have higher hazards than women. That is why we designate x_1 (indicator for sex of the patient) as “protected” here and look for the “best” covariates to keep among $z_1, \dots, z_6$. We develop selection methods that choose different subsets of the z_j ’s for different purposes, as opposed to being content with one subset to serve all purposes.

1.2 Some Selection Methods

There are rather few well-developed variable selection methods for the Cox model. Methods involving pretesting of coefficients and variants of backward and forward regression can be put forward, in partial analogy with linear or generalized linear regression theory. The general model averaging theory that we develop in Section 8 actually accurately describes the performance of such methods. Fan and Li (2002) proposed a penalized version of the log-partial likelihood, with a penalty called the smoothly clipped absolute deviation (SCAD). This SCAD penalty depends on two unknown parameters, one fixed at a predetermined value and the other chosen through an approximation to generalized cross-validation. This approach was further extended by Cai, Fan, Li, and Zhou (2005) for marginal hazards models. In that article, different regression coefficients were allowed to have different penalties and regularization parameters. Tibshirani (1997) used the lasso method for variable selection in the Cox model; this and similar L_1 -based methods refined later by Efron, Hastie, Johnstone, and Tibshirani (2004) are of particular value when the number q of nonprotected covariates is large. Bunea and McKeague (2005) introduced a penalized likelihood, where now the penalty depends on both the number of parameters in the parametric part of the model and on the

Nils Lid Hjort is Professor, Department of Mathematics, University of Oslo, N-0316 Oslo, Norway (E-mail: nils@math.uio.no). Gerda Claeskens is Associate Professor, ORSTAT and University Centre for Statistics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium (E-mail: gerda.claeskens@econ.kuleuven.be). The article has benefitted from constructive suggestions from the associate editor, editor Frank Samaniego, and three anonymous reviewers. The authors gratefully acknowledge hospitality and partial funding from the Department of Mathematics at the University of Oslo, the Norwegian Research Council via its BeMatA programme, the Institut de Statistique at Université Catholique de Louvain, and the University Centre for Statistics at Katholieke Universiteit Leuven at reciprocal visits.

number of components in the sieve construction to estimate the unknown baseline hazard function.

More traditional model selection methods, such as the Akaike information criterion (AIC) and Bayes information criterion (BIC), are not automatically defined for Cox models, because there is no workable full likelihood for data. One may, however, choose to use the partial likelihoods, say $L_{n,S}$ for the model that only uses covariates $z_{i,j}$ for $j \in S$, which leads to

$$\begin{aligned} \text{AIC}_{n,S} &= 2 \log L_{n,S}(\hat{\beta}_S, \hat{\gamma}_S) - 2(p + |S|), \\ \text{BIC}_{n,S} &= 2 \log L_{n,S}(\hat{\beta}_S, \hat{\gamma}_S) - (p + |S|) \log n, \end{aligned} \quad (2)$$

in terms of the Cox estimators $(\hat{\beta}_S, \hat{\gamma}_S)$ inside the S submodel. Here $|S|$ denotes the number of elements in S , and the model with the highest score is selected. These model selection schemes are easily implemented using software for handling the Cox regression model. Volinsky and Raftery (2000) investigated some aspects of the BIC scheme, including discussion of other penalty factors, along with versions of Bayesian model-averaging strategies for the Cox model. Less theory has been developed for the AIC and BIC methods valid for the Cox model than for fully parametric regression models, however. It should be noted that these criteria work only with the parametric part of the (1) model, thus ignoring the nonparametric part. Therefore, it is not clear whether model selectors using (2) are relevant when it comes to consequences for questions that relate to all of the (1) model, like survival probabilities and median survival time.

1.3 Focused Information Criteria and Model Averaging

Different selection methods find different “best” subsets, as is dramatically shown in the Danish melanoma dataset with the AIC and the BIC (see Sec. 9.2). What all of the aforementioned methods have in common is that they advocate one and only one final model, regardless of its intended use, whether this involves predicting median survival time or estimating survival chances for patients with unusual characteristics and so on. We develop an FIC that, for each parameter of interest, finds the best submodel for that purpose. Specifically, for a given focus parameter $\mu(\beta, \gamma, H_0)$, where H_0 is the cumulative baseline hazard rate, we are able to estimate the mean squared error (MSE) for each of the many candidate estimators. The FIC strategy is to select the model with lowest possible MSE estimate. We do not view this as a paradox, even if it means leaving behind the traditionally strong and conceptually sirenning paradigm of finding one adequate model to explain all aspects and facets of data. See also Longford (2005).

When a model selection scheme like the AIC or FIC is followed to produce an estimator, it is important to realize that the real variance involved is larger than it would be had the selected model been given in advance. Studying statistical properties of such postselection estimators involves more work than simply understanding the limit distributions of the Cox estimators. In our work we reach precise large-sample results for a broad class of “compromise estimators” that interpolate between all candidate models, with the postselection estimators constituting special cases. The limit distributions involved are not normal, but rather nonlinear mixtures of different normals.

The theory and results for our FIC and model average estimators parallel the development and findings in our earlier articles (Claeskens and Hjort 2003; Hjort and Claeskens 2003a,b, hereafter referred to as CH and HC 2003). These articles were concerned with likelihood methods for general parametric models, including various regression models, but the results reached there do not capture and cannot be applied directly inside the Cox regression model. This is partly because of the censoring and of the semiparametric nature of model (1), rendering analysis of estimators that combine both H_0 and (β, γ) estimators difficult. Thus a separate development for building a proper FIC for the Cox model is needed, supplemented with proper model averaging methods.

While working on this development, we have learned that the CH and HC (2003) theory and methods carry over with reasonable ease to situations that involve only the regression parameters (β, γ) ; in other words, as long as questions are posed that can be answered in terms of (β, γ) , we do not need significant extensions of the already available theory. This comment will be seen to apply to results for the AIC strategy as well. Many questions of interest relate to the full model (1), however, including the hazard rate part, like the median survival time $H_0^{-1}(\log 2 / \exp(\mathbf{x}^t \beta + \mathbf{z}^t \gamma))$, the survival probability $\exp\{-\exp(\mathbf{x}^t \beta + \mathbf{z}^t \gamma) H_0(t)\}$, and conditional survival probabilities given that one has survived up to a certain time point. It is for such reasons that the CH and HC (2003) theory needs more work to appropriately extend it, as we discuss later.

1.4 The Present Article

The article is organized as follows. In Section 2 we set up the basic framework, properly defining all subset estimators $\hat{\mu}_S$ for a given focus parameter $\mu(H_0, \beta, \gamma)$, and providing a local neighborhood formulation that turns out to give fruitful large-sample approximations to modeling bias, variance, and distributions of estimators. In Sections 3 and 4 we develop theory for describing the behavior of submodel-based estimators $\hat{\beta}_S$ and $\hat{\gamma}_S$ for the regression coefficients and $\hat{H}_{0,S}$ for the cumulative hazard. We use this in Section 5 to provide precise large-sample results for limit distributions and limiting risk for all submodel estimators. This is a harder task than proving limit theorems for the Cox estimators, in that these also need to be studied outside model conditions and because we need to consider simultaneous aspects of all of the estimators involved. In Section 6 we go through a list of particularly important parameters of interest, including survival probability curves for given strata of patients, median survival times, and relative risks. We also describe how to estimate various quantities necessary for implementing the FIC methods. In Section 7 we give the proper machinery for the FIC and its averaged versions. Then in Section 8 we provide a “master theorem” that accurately describes the limit distribution for a large class of model averaging estimators. In particular, this provides precise descriptions of the large-sample behavior of all postselection strategies, like the AIC and the FIC. Also in Section 8 is a discussion of unfortunate side consequences for consistent model selection methods, like the BIC; the maximal risk using, for example, the BIC, divided by the maximal risk using, for example, the FIC, tends to infinity with growing n . In Section 9 we illustrate our methods in some settings with simulated data and analyze the Danish melanoma survival dataset. We end the article with a list of

concluding remarks in Section 10, some of which might lead to further research work. We provide proofs of the lemmas and theorem in the Appendix.

2. A FRAMEWORK FOR COVARIATE SUBSET SELECTION

Working inside the (1) model, with lifetime data observed or partly observed over a time horizon $[0, \tau]$, the log-partial likelihood can be written as

$$\begin{aligned} \log L_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \sum_{i=1}^n \int_0^\tau \left[\mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\gamma} \right. \\ &\quad \left. - \log \left\{ \sum_{i=1}^n Y_i(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\gamma}) \right\} \right] dN_i(u), \quad (3) \end{aligned}$$

where $Y_i(u) = I\{t_i \geq u\}$ and $dN_i(u) = I\{t_i \in [u, u + du], \delta_i = 1\}$. The maximum partial likelihood estimators, also referred to as the Cox estimators, are the $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ values maximizing (3). The theory developed in this article partly uses the large-sample theory associated with counting processes and martingales, as expositied by, for example, Andersen et al. (1993), where results are most comfortably reached if the upper time limit τ is finite, so we assume it to be so; somewhat more technical assumptions would be needed if we wished to obtain results valid for $\tau = \infty$.

For each subset S of $\{1, \dots, q\}$, we may study the model indexed by $(\boldsymbol{\beta}, \boldsymbol{\gamma}_S)$, where $\boldsymbol{\gamma}_S$ contains precisely those γ_j coefficients where $j \in S$, that is, corresponding to including in the model only those $z_{i,j}$'s where $j \in S$ and excluding those with $j \notin S$. There are a total of 2^q such submodels to consider. Sometimes several of these may be ruled out on a priori grounds, for example, when there is a natural ordering in complexity, in which case only the $q + 1$ nested submodels of $\{1, \dots, q\}$ are considered. For each submodel S there are Cox estimators $(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S)$, and also an accompanying Aalen–Breslow type estimator of the cumulative baseline hazard function, namely

$$\hat{H}_{0,S}(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n Y_i(u) \exp(\mathbf{x}_i^t \hat{\boldsymbol{\beta}}_S + \mathbf{z}_{i,S}^t \hat{\boldsymbol{\gamma}}_S)}, \quad (4)$$

where $\mathbf{z}_{i,S}$ represents the components $z_{i,j}$ of $\mathbf{z}_i$ for which $j \in S$. Accordingly, for given estimand of interest, say $\mu = \mu(\boldsymbol{\beta}, \boldsymbol{\gamma}, H_0)$, there is a list of potential estimators,

$$\hat{\mu}_S = \mu(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S, \mathbf{0}_{S^c}, \hat{H}_{0,S}), \quad (5)$$

one for each submodel. The notation indicates that we use the null value $\gamma_j = 0$ for $j \notin S$, that is, for $j \in S^c$, the complement set. The μ in question may also depend on covariate positions $\mathbf{x}$ and $\mathbf{z}$, as exemplified in Sections 6 and 9.

We study questions of covariate inclusion and exclusion inside a large-sample framework where $\boldsymbol{\gamma}$ is small or moderate and where the largest of the models, the one containing all $p + q$ covariates, contains the truth. More specifically, the real hazard rate functions are taken to be

$$h_{i,\text{true}}(u) = h_0(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\gamma} / \sqrt{n}) \quad \text{for } i = 1, \dots, n, \quad (6)$$

for suitable $(\beta_1, \dots, \beta_p)^t$ and $(\eta_1, \dots, \eta_q)^t$ not depending on n . This turns out to be a fruitful framework for deriving accurate approximations to modeling bias and variances and hence MSEs for different estimators, essentially because variances and squared biases now become exchangeable currencies, both of order $O(1/n)$.

The assumption (6) is a crucial one for deriving our results and for the new model selection proposals, and warrants a brief and hopefully clarifying discussion before we go further. Its chief nonstandard feature compared with (1) is that we let $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(n)} = \boldsymbol{\eta} / \sqrt{n}$ depend on the sample size. This is a mathematical construction, chosen with the aim of deriving informative limit distributions and good approximations for MSEs of estimators, for example. Thus (6) is not meant to convey that we interpret the $\boldsymbol{\gamma}$ of (1) to change value from $n = 200$ to $n = 201$, for example; this would also clash with Kolmogorov consistency. The trick to deriving results inside the (6) framework is akin to using local alternatives to derive approximations to local power of hypothesis tests. Thus our limit distribution results, arrived at in the following sections, are really obtained in a “triangular” framework, with the true model depending on n . To avoid overburdening the notation, we omit the n in $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(n)}$ and so on, however. (See also the general discussion surrounding these issues in CH 2003 and HC 2003a,b.)

3. SUBMODEL ESTIMATORS FOR HAZARD REGRESSION COEFFICIENTS

In this section we develop theory for the large-sample behavior of all submodel estimators $(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S)$. This requires some modifications and extensions of standard theory for the Cox regression models, in that we need to analyze estimators for models that are perhaps approximately but not fully correct. We also need an apparatus for handling estimators from different submodels simultaneously. (See Andersen and Gill 1982; Gill 1984; and Andersen et al. 1993 for such “standard theory.”)

To properly analyze the submodel estimators, we need to introduce certain random quantities and their limit functions. Let

$$\mathbf{G}_n^{(0)}(u, \boldsymbol{\beta}, \boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^n Y_i(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\gamma}),$$

$$\mathbf{G}_n^{(1)}(u, \boldsymbol{\beta}, \boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^n Y_i(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\gamma}) \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_i \end{pmatrix},$$

$$\mathbf{G}_n^{(2)}(u, \boldsymbol{\beta}, \boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^n Y_i(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\gamma}) \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_i \end{pmatrix} \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_i \end{pmatrix}^t,$$

and $\mathbf{E}_n(u, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbf{G}_n^{(1)}(u, \boldsymbol{\beta}, \boldsymbol{\gamma}) / \mathbf{G}_n^{(0)}(u, \boldsymbol{\beta}, \boldsymbol{\gamma})$. We also need the subfunctions associated with using the subset S instead of the full $\{1, \dots, q\}$; thus $\mathbf{G}_{n,S}^{(1)}$ is the $(p + |S|)$ -vector where the first p components make up $\mathbf{G}_{n,0}^{(1)}$ and the next $|S|$ components define $\mathbf{G}_{n,1,S}^{(1)}$, and similarly with the ratio $\mathbf{E}_{n,S}$, which has p components giving $\mathbf{E}_{n,0}$ and then $|S|$ components defining $\mathbf{E}_{n,1,S}$.

As is commonly assumed in treatises on the Cox regression model, we postulate that these functions have limits in probability $g^{(0)}(s, \boldsymbol{\beta}, \boldsymbol{\gamma})$, $\mathbf{g}^{(1)}(s, \boldsymbol{\beta}, \boldsymbol{\gamma})$, and $\mathbf{g}^{(2)}(s, \boldsymbol{\beta}, \boldsymbol{\gamma})$, and that these limit functions are continuous in s . Actually, because we

work under the assumption (6), we are more concerned with the related condition that $G_n^{(0)}(s, \boldsymbol{\beta}, \boldsymbol{\eta}/\sqrt{n}) \rightarrow_p g^{(0)}(s, \boldsymbol{\beta}, \mathbf{0})$, and so on. We also write $\mathbf{e}(s, \boldsymbol{\beta}, \mathbf{0})$ for the limit function of $G_n^{(1)}(s, \boldsymbol{\beta}, \boldsymbol{\eta}/\sqrt{n})/G_n^{(0)}(s, \boldsymbol{\beta}, \boldsymbol{\eta}/\sqrt{n})$.

Let $\mathbf{U}_n$ and $\mathbf{V}_n$ be the derivatives with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ of the log-likelihood normalized by n^{-1} , so that

$$\begin{pmatrix} \mathbf{U}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) \\ \mathbf{V}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) \end{pmatrix} = n^{-1} \sum_{i=1}^n \int_0^\tau \left\{ \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_i \end{pmatrix} - \mathbf{E}_n(u, \boldsymbol{\beta}, \boldsymbol{\gamma}) \right\} dN_i(u). \quad (7)$$

Also let $\mathbf{I}_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$ be the $(p+q) \times (p+q)$ matrix of second-order derivatives, leading to

$$-\mathbf{I}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \int_0^\tau \boldsymbol{\Sigma}_n(u, \boldsymbol{\beta}, \boldsymbol{\gamma}) G_n^{(0)}(u, \boldsymbol{\beta}, \boldsymbol{\gamma}) h_0(u) du,$$

in which $\boldsymbol{\Sigma}_n = \mathbf{G}_n^{(2)}/G_n^{(0)} - \mathbf{E}_n \mathbf{E}_n^t$.

Under standard assumptions about the covariate sequences $\mathbf{x}_i$ and $\mathbf{z}_i$, and in the framework defined by (6), it follows that $-\mathbf{I}_n(\boldsymbol{\beta}, \boldsymbol{\eta}/\sqrt{n})$ as well as $-\mathbf{I}_n(\boldsymbol{\beta}, \mathbf{0})$ have as a limit in probability a $(p+q) \times (p+q)$ matrix,

$$\mathbf{J}_{\text{full}} = \begin{pmatrix} \mathbf{J}_{00} & \mathbf{J}_{01} \\ \mathbf{J}_{10} & \mathbf{J}_{11} \end{pmatrix}, \quad (8)$$

which we also take to be positive definite (see Andersen et al. 1993 for more details). In fact, $\mathbf{J}_{n,\text{full}} = -\mathbf{I}_n(\hat{\boldsymbol{\beta}}_{\text{full}}, \hat{\boldsymbol{\gamma}}_{\text{full}})$ also tends in probability to $\mathbf{J}_{\text{full}}$, under condition (6). The estimator that we use for $\mathbf{J}_{\text{full}}$ is

$$\begin{aligned} \hat{\mathbf{J}}_{\text{full}} &= \int_0^\tau \boldsymbol{\Sigma}_n(u, \hat{\boldsymbol{\beta}}_{\text{full}}, \hat{\boldsymbol{\gamma}}_{\text{full}}) G_n^{(0)}(u, \hat{\boldsymbol{\beta}}_{\text{full}}, \hat{\boldsymbol{\gamma}}_{\text{full}}) d\hat{H}_{0,\text{full}}(u) \\ &= \int_0^\tau \boldsymbol{\Sigma}_n(u, \hat{\boldsymbol{\beta}}_{\text{full}}, \hat{\boldsymbol{\gamma}}_{\text{full}}) n^{-1} \sum_{i=1}^n dN_i(u) \\ &= n^{-1} \sum_{i=1}^n \boldsymbol{\Sigma}_n(t_i, \hat{\boldsymbol{\beta}}_{\text{full}}, \hat{\boldsymbol{\gamma}}_{\text{full}}) \delta_i. \end{aligned} \quad (9)$$

We shall also have occasion to need the $(p+|S|) \times (p+|S|)$ submatrix $\mathbf{J}_S$, with blocks, say, $\mathbf{J}_{00}, \mathbf{J}_{01,S}, \mathbf{J}_{10,S}$, and $\mathbf{J}_{11,S}$. It is convenient to phrase some of the results in terms of the projection function $\boldsymbol{\pi}_S: \mathcal{R}^q \rightarrow \mathcal{R}^{|S|}$, which takes $\mathbf{v} = (v_1, \dots, v_q)^t$ to its subvector $\mathbf{v}_S$ with those v_j for which $j \in S$. Thus $\boldsymbol{\pi}_S$ is an $|S| \times q$ matrix of 1's and 0's.

The following key lemma (the proof of which is given in the App.) gives the precise large-sample behavior of the S -submodel Cox estimators, under the assumption (6). We assume that "ordinary regularity conditions," as spelled out and discussed by Andersen et al. (1993, chap. VII) are in force; these may actually also be substantially weakened, as discussed by Hjort (1992) and Hjort and Pollard (1996).

Lemma 1. Assume that the conditions just described are in force. Then, under the sequence of true hazard rate functions (6),

$$\begin{pmatrix} \sqrt{n}(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}) \\ \sqrt{n}\hat{\boldsymbol{\gamma}}_S \end{pmatrix} \rightarrow_d \begin{pmatrix} \mathbf{B}_S \\ \mathbf{C}_S \end{pmatrix} \sim N_{p+|S|} \left(\mathbf{J}_S^{-1} \begin{pmatrix} \mathbf{J}_{01} \\ \boldsymbol{\pi}_S \mathbf{J}_{11} \end{pmatrix} \boldsymbol{\eta}, \mathbf{J}_S^{-1} \right).$$

Armed with this lemma, we may derive useful expressions for the approximate MSE of estimators $\mu(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S, \mathbf{0}_{S^c})$ of estimands of the type $\mu(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Because we take an interest in

more general estimands, which also may depend on H_0 , further efforts are needed to determine the behavior of the $\hat{H}_{0,S}$ estimators.

4. SUBMODEL ESTIMATORS FOR THE CUMULATIVE BASELINE HAZARD

Inside a submodel S , which gives maximum partial likelihood estimators $(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S)$ for the model (3), we now study the accompanying Aalen–Breslow type estimator given in (4) for the cumulative hazard function $H_0(t) = \int_0^t h_0(u) du$. To reach a precise result, first consider

$$W_n(t) = n^{-1/2} \int_0^t \frac{\sum_{i=1}^n dM_i(u)}{G_n^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})}.$$

This is a martingale with variance function converging toward $\int_0^t g^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})^{-1} dH_0(u)$, which implies that the $W_n(\cdot)$ process tends in distribution to a Gaussian mean-0 martingale $W(\cdot)$ with $\text{var} dW(u) = dH_0(u)/g^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})$. We also find that the W_n process and the vector of $\sqrt{n}\mathbf{U}_n(\boldsymbol{\beta}, \mathbf{0})$ and $\sqrt{n}\mathbf{V}_n(u, \boldsymbol{\beta}, \mathbf{0})$ are independent in the limit; that is, the W process becomes independent of each $\mathbf{B}_S$ and $\mathbf{C}_S$ of Lemma 1. To see this, work first with $\text{cov}\{\sqrt{n}\mathbf{U}_n(\boldsymbol{\beta}, \mathbf{0}), dW_n(u)\}$, which, by martingale theory, can be expressed as the mean of $\mathbf{S}_n$, say, where

$$\mathbf{S}_n = n^{-1} \sum_{i=1}^n \{\mathbf{x}_i - \mathbf{E}_{n,0}(u, \boldsymbol{\beta}, \mathbf{0})\} \frac{Y_i(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\eta}/\sqrt{n})}{G_n^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})},$$

which is seen to be composed of $\mathbf{G}_{n,0}^{(1)}(u, \boldsymbol{\beta}, \mathbf{0}) - \mathbf{E}_{n,0}(u, \boldsymbol{\beta}, \mathbf{0}) \times G_n^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})$, which vanishes, plus a term of order $O_p(n^{-1/2})$. A similar calculation confirms the claim for $\sqrt{n}\mathbf{V}_n(u, \boldsymbol{\beta}, \mathbf{0})$ and W_n .

For the next central result (the proof of which is given in the App.), we introduce the $(p+q)$ -vector function

$$\mathbf{F}(t) = \int_0^t \mathbf{e}(u, \boldsymbol{\beta}, \mathbf{0}) dH_0(u) = \begin{pmatrix} \mathbf{F}_0(t) \\ \mathbf{F}_1(t) \end{pmatrix},$$

where the first p components compose $\mathbf{F}_0(t)$ and the final q components make up $\mathbf{F}_1(t)$. We also use $\mathbf{F}_{1,S}(t)$ to denote the subset of $\mathbf{F}_1(t)$ with components belonging to subset S , and finally $\mathbf{F}_S(t)$ for the $(p+|S|)$ -vector with $\mathbf{F}_0(t)$ and $\mathbf{F}_{1,S}(t)$.

Lemma 2. Under the assumption (6), in concert with other conditions stated in connection with Lemma 1, the $A_{n,S}(t) = n^{1/2}\{\hat{H}_{0,S}(t) - H_0(t)\}$ process tends in distribution to the process

$$A_S(t) = W(t) - \begin{pmatrix} \mathbf{F}_0(t) \\ \mathbf{F}_{1,S}(t) \end{pmatrix}^t \begin{pmatrix} \mathbf{B}_S \\ \mathbf{C}_S \end{pmatrix} + \mathbf{F}_1(t)^t \boldsymbol{\eta}.$$

Note that $\text{var} A_S(t) = \int_0^t g^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})^{-1} dH_0(u) + \mathbf{F}_S(t)^t \times \mathbf{J}_S^{-1} \mathbf{F}_S(t)$, growing larger when more covariates are included. This needs to be weighted against its bias level, which can be determined from Lemmas 1 and 2. An expression for the bias will also flow from the work of the next section.

5. LIMITING RISK OF SUBMODEL ESTIMATORS

Consider an estimand of the general type $\mu(\boldsymbol{\beta}, \boldsymbol{\gamma}, H_0(t))$, with t kept fixed for the moment. There is one potential estimator, $\hat{\mu}_S = \mu(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S, \mathbf{0}_{S^c}, \hat{H}_{0,S}(t))$, for each regressor sub-

set $S \subset \{1, \dots, q\}$. We derive a precise limit distribution result for $\hat{\mu}_S$. Our limiting risk results involve concise expressions for bias and variance in terms of the quantities

$$\begin{aligned}\Omega_S &= \pi_S^t \mathbf{K}_S \pi_S \mathbf{K}^{-1}, \\ \omega &= \mathbf{J}_{10} \mathbf{J}_{00}^{-1} \frac{\partial \mu}{\partial \beta} - \frac{\partial \mu}{\partial \gamma}, \quad \text{and} \\ \kappa &= \kappa(t) = \{\mathbf{J}_{10} \mathbf{J}_{00}^{-1} \mathbf{F}_0(t) - \mathbf{F}_1(t)\} \frac{\partial \mu}{\partial H_0}.\end{aligned}\quad (10)$$

Here $\mathbf{K} = \mathbf{J}^{11}$ is a $q \times q$ matrix, computed from the inverse of $\mathbf{J}_{\text{full}}$, and, similarly,

$$\mathbf{K}_S = \mathbf{J}^{11,S} = (\pi_S \mathbf{K}^{-1} \pi_S^t)^{-1}$$

is the lower-right corner $|S| \times |S|$ submatrix of the inverse of $\mathbf{J}_S$. The partial derivatives are evaluated at the center point $(\beta, \mathbf{0}, H_0(t))$; thus both $\frac{\partial \mu}{\partial H_0}$ and $\kappa = \kappa(t)$ depend on the t under consideration. Note that both ω and κ are of dimension q . Finally, define

$$\begin{aligned}\tau_0^2 &= \left(\frac{\partial \mu}{\partial H_0} \right)^2 \int_0^t \frac{dH_0(u)}{g^{(0)}(u, \beta, \mathbf{0})} \\ &+ \left\{ \frac{\partial \mu}{\partial \beta} - \frac{\partial \mu}{\partial H_0} \mathbf{F}_0(t) \right\}^t \mathbf{J}_{00}^{-1} \left\{ \frac{\partial \mu}{\partial \beta} - \frac{\partial \mu}{\partial H_0} \mathbf{F}_0(t) \right\},\end{aligned}\quad (11)$$

which is the minimal possible limiting variance of the $\hat{\mu}_S$ estimators. The model underlying the data is again taken to be that of (6), under which $\mu_{\text{true}} = \mu(\beta, \eta/\sqrt{n}, H_0(t))$.

Lemma 3. Assume that the μ estimand has continuous derivatives in a neighborhood of $(\beta, \mathbf{0}, H_0(t))$. Under the conditions laid out for Lemmas 1 and 2, and under the circumstances in (6), the variable $\Lambda_{n,S} = \sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ tends in distribution to

$$\Lambda_S = \left(\frac{\partial \mu}{\partial \beta} \right)^t \mathbf{B}_S + \left(\frac{\partial \mu}{\partial \gamma} \right)^t \mathbf{C}_S - \left(\frac{\partial \mu}{\partial \gamma} \right)^t \eta + \frac{\partial \mu}{\partial H_0} A_S(t).\quad (12)$$

This is a normal variable with mean and variance equal to $(\omega - \kappa)^t (\mathbf{I} - \Omega_S) \eta$ and $\tau_0^2 + (\omega - \kappa)^t \Omega_S \mathbf{K} \Omega_S^t (\omega - \kappa)$.

Following the details of the proof, given in the Appendix, we also establish a quite fruitful representation of the limit distribution, namely

$$\begin{aligned}\Lambda_S &= \left(\frac{\partial \mu}{\partial H_0} \right) W(t) + \left(\frac{\partial \mu}{\partial \beta} - \frac{\partial \mu}{\partial H_0} \mathbf{F}_0(t) \right)^t \mathbf{J}_{00}^{-1} \mathbf{U} \\ &+ (\omega - \kappa)^t \{(\mathbf{I} - \Omega_S) \eta - \Omega_S \mathbf{V}'\},\end{aligned}\quad (13)$$

in which $\mathbf{U}$ and $\mathbf{V}'$ are independent and $N_p(\mathbf{0}, \mathbf{J}_{00})$ and $N_q(\mathbf{0}, \mathbf{K})$. It also follows from this that the MSE of the large-sample limit of n times $\hat{\mu}_S$ (i.e., the limiting risk associated with using the S subset) is

$$\begin{aligned}\text{E}\Lambda_S^2 &= \tau_0^2(t) + (\omega - \kappa(t))^t \{(\mathbf{I} - \Omega_S) \eta \eta^t (\mathbf{I} - \Omega_S^t) + \Omega_S \mathbf{K} \Omega_S^t\} \\ &\times (\omega - \kappa(t)),\end{aligned}\quad (14)$$

allowing our notation here to reflect that both τ_0 and κ depend on the t engaged in the estimand $\mu = \mu(\beta, \gamma, H_0(t))$.

Remark 1. There is a result corresponding to that of lemma 3 in CH and HC (2003), valid for general parametric families but involving only a quantity similar to the ω . It is the semipara-

metric nature of the Cox regression model that here leads to the more general $\omega - \kappa$ quantity.

Remark 2. If $\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d \Lambda$, say, then we use “limit risk” in the sense of the MSE $\text{E}\Lambda^2$ of the limit distribution. This is typically (but not always) the same as the limit of n times the MSE of $\hat{\theta}$. The limit risk may also be defined as the limit of $r(c) = \lim_n \text{E} \min\{n(\hat{\theta} - \theta)^2, c\}$ as $c \rightarrow \infty$.

6. RISK CALCULATION AND ESTIMATION FOR IMPORTANT ESTIMANDS

Note that the limit distributions and limiting risks derived in the previous section depend crucially on both $\tau_0(t)$ and the coefficients of $\omega - \kappa(t)$, which vary from one parameter to the next. We now illustrate this for a brief list of examples, before turning to the task of estimating these and other quantities involved in the limiting risk expressions. The important case of the median survival time or, more generally, the task of estimating the quantile distribution of the survival time, needs some technical development of separate interest, which we provide in Section 6.2.

6.1 A List of Foci

1. One may naturally compare hazard level for individuals with covariate $\mathbf{x}$ with hazard level for those with covariate $\mathbf{x}_0$ using a hazard ratio of, say, $h(s | \mathbf{x}, \mathbf{z})/h(s | \mathbf{x}_0, \mathbf{z}) = \exp\{(\mathbf{x} - \mathbf{x}_0)^t \beta\}$. This focus parameter has $\omega = \exp\{(\mathbf{x} - \mathbf{x}_0)^t \beta\} \mathbf{J}_{10} \mathbf{J}_{00}^{-1} (\mathbf{x} - \mathbf{x}_0)$ and $\kappa = \mathbf{0}$.

2. A natural parameter of interest is the relative risk $\mu = \exp(\mathbf{x}^t \beta + \mathbf{z}^t \gamma)$ at position $(\mathbf{x}, \mathbf{z})$ in the covariate space; here $\omega = \exp(\mathbf{x}^t \beta) (\mathbf{J}_{10} \mathbf{J}_{00}^{-1} \mathbf{x} - \mathbf{z})$, whereas $\kappa = \mathbf{0}$. The quantity just discussed can be viewed as the relative risk in comparison with an individual with covariates $(\mathbf{x}, \mathbf{z}) = (\mathbf{0}, \mathbf{0})$. This is a natural quantity in situations where the covariates have been centered to have mean 0; in this case the “relative” in “relative risk” would mean in comparison with the “average” individual. Similarly, if $\mathbf{x}$ and $\mathbf{z}$ represent risk factors, scaled such that the zero level corresponds to normal healthy conditions and positive values correspond to increased risk, then the $\mu = \mu(\mathbf{x}, \mathbf{z})$ represents the relative risk increase at level $(\mathbf{x}, \mathbf{z})$ in comparison with normal health level. In yet other situations it would be more natural to compare individuals with an existing or hypothesized individual with suitable null covariates $(\mathbf{x}_0, \mathbf{z}_0)$, say. This corresponds to focusing on the relative risk $\mu = \exp\{(\mathbf{x} - \mathbf{x}_0)^t \beta + (\mathbf{z} - \mathbf{z}_0)^t \gamma\}$ and leads to $\kappa = \mathbf{0}$ and

$$\omega = \exp\{(\mathbf{x} - \mathbf{x}_0)^t \beta\} \{ \mathbf{J}_{10} \mathbf{J}_{00}^{-1} (\mathbf{x} - \mathbf{x}_0) - (\mathbf{z} - \mathbf{z}_0) \}.\quad (15)$$

Note in particular that different covariate levels give different ω vectors, which, in view of Lemma 3 and risk expression (14), means that there might well be different optimal S submodels for different covariate regions. This is accounted for in our FIC for model selection, as discussed in Section 7.

3. For the problem of estimating $H_0(t)$ separately, the ω vector is $\mathbf{0}$, whereas $\kappa = \mathbf{J}_{10} \mathbf{J}_{00}^{-1} \mathbf{F}_0(t) - \mathbf{F}_1(t)$.

4. Estimating a survival probability for a given individual translates to

$$\text{Su}(t | \mathbf{x}, \mathbf{z}) = \exp\{-\exp(\mathbf{x}^t \beta + \mathbf{z}^t \gamma) H_0(t)\},$$

for which we find that $\omega = -\text{Su}(t | \mathbf{x}, \mathbf{z}) H_0(t) (\mathbf{J}_{10} \mathbf{J}_{00}^{-1} \mathbf{x} - \mathbf{z})$ and $\kappa = -\text{Su}(t | \mathbf{x}, \mathbf{z}) \exp(\mathbf{x}^t \beta) \{ \mathbf{J}_{10} \mathbf{J}_{00}^{-1} \mathbf{F}_0(t) - \mathbf{F}_1(t) \}$.

5. Consider now a patient's chance of surviving t given that he or she has managed to survive up to time t_0 . This probability is $\exp[-\{H_0(t) - H_0(s)\} \exp(\mathbf{x}^t \boldsymbol{\beta} + \mathbf{z}^t \boldsymbol{\gamma})]$. Handling this estimand calls for some modifications of Lemmas 2 and 3, in that $\int_{t_0}^t dH_0(u)$ is at work rather than the full $H_0(t)$. Lemma 2 may be extended to reach parallel results involving $A_S(t) - A_S(t_0)$ rather than simply $A_S(t)$ without serious difficulties. This includes revised definitions of κ and τ_0^2 , replacing $\mathbf{F}_0(t)$ and $\mathbf{F}_1(t)$ with $\mathbf{F}_0(t) - \mathbf{F}_0(t_0)$ and $\mathbf{F}_1(t) - \mathbf{F}_1(t_0)$.

6.2 Estimating Median Survival Time

A patient's median survival time, in terms of his or her covariates, can be expressed as $\xi = H_0^{-1}(\log 2 / \exp(\mathbf{x}^t \boldsymbol{\beta} + \mathbf{z}^t \boldsymbol{\gamma}))$. That this is a quantity of serious interest, and sometimes more important than, say, the mean survival time, was made clear by, for example, Gould (1995). Earlier work on conditional median survival time includes that of Dabrowska and Doksum (1987) and Burr and Doss (1993). Handling the case of such conditional quantiles here, in general,

$$\xi(r) = H_0^{-1} \left(-\frac{\log(1-r)}{\exp(\mathbf{x}^t \boldsymbol{\beta} + \mathbf{z}^t \boldsymbol{\gamma})} \right) \quad \text{for } 0 < r < 1,$$

requires some separate development, and it is not a priori clear that the limiting distribution of, say,

$$\sqrt{n}(\hat{\xi}_S - \xi_{\text{true}}) = \sqrt{n} \left\{ \hat{H}_{0,S}^{-1} \left(\frac{\log 2}{\exp(\mathbf{x}^t \hat{\boldsymbol{\beta}}_S + \mathbf{z}^t \hat{\boldsymbol{\gamma}}_S)} \right) - H_0^{-1} \left(\frac{\log 2}{\exp(\mathbf{x}^t \boldsymbol{\beta} + \mathbf{z}^t \boldsymbol{\gamma} / \sqrt{n})} \right) \right\},$$

has the same appealing structure as in Lemma 3, because the $\xi(\boldsymbol{\beta}, \boldsymbol{\gamma}, H_0)$ under consideration now does not depend on H_0 only at a single value.

Consider an estimand of the general form $\xi = H_0^{-1}(f(\boldsymbol{\beta}, \boldsymbol{\gamma}))$, where $f(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is some smooth function of the regression coefficients and for which we contemplate using any of the estimates

$$\begin{aligned} \hat{\xi}_S &= \hat{H}_{0,S}^{-1}(f(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S, \mathbf{0}_S)) \\ &= \sup\{t: \hat{H}_{0,S}(t) \leq f(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S, \mathbf{0}_S)\}. \end{aligned}$$

The following result is proved in the Appendix.

Lemma 4. Assume that conditions laid out for Lemmas 1 and 2 are in force and that $f(\boldsymbol{\beta}, \boldsymbol{\gamma})$ has continuous derivatives in a neighborhood of $(\boldsymbol{\beta}, \mathbf{0})$. Then, under the circumstances in (6), the variable $\Lambda_{n,S} = \sqrt{n}(\hat{\xi}_S - \xi_{\text{true}})$ tends in distribution to

$$\Lambda_S = h_0(\xi_0)^{-1} \{-W(\xi_0) + \mathbf{R}_0^t \mathbf{J}_{00}^{-1} \mathbf{U} + \boldsymbol{\zeta}^t \boldsymbol{\Omega}_S \mathbf{V}' - \boldsymbol{\zeta}^t (\mathbf{I} - \boldsymbol{\Omega}_S) \boldsymbol{\eta}\},$$

with $\mathbf{U}$ and $\mathbf{V}'$ as in (13), where $\xi_0 = H_0^{-1}(f(\boldsymbol{\beta}, \mathbf{0}))$, $\mathbf{R}_0 = \mathbf{F}_0(\xi_0) + \frac{\partial f}{\partial \boldsymbol{\beta}}$, $\mathbf{R}_1 = \mathbf{F}_1(\xi_0) + \frac{\partial f}{\partial \boldsymbol{\gamma}}$, and $\boldsymbol{\zeta} = \mathbf{R}_1 - \mathbf{J}_{10} \mathbf{J}_{00}^{-1} \mathbf{R}_0$ and where the partial derivatives of f are evaluated at $(\boldsymbol{\beta}, \mathbf{0})$. The limit distribution is normal, with mean and variance equal to

$$\begin{aligned} &-h_0(\xi_0)^{-1} \boldsymbol{\zeta}^t (\mathbf{I} - \boldsymbol{\Omega}_S) \boldsymbol{\eta} \quad \text{and} \\ &h_0(\xi_0)^{-2} \{\text{var } W(\xi_0) + \mathbf{R}_0^t \mathbf{J}_{00}^{-1} \mathbf{R}_0 + \boldsymbol{\zeta}^t \boldsymbol{\Omega}_S \mathbf{K} \boldsymbol{\Omega}_S^t \boldsymbol{\zeta}\}. \end{aligned}$$

The limit distribution involves the baseline hazard rate $h_0(u)$, which for its estimation would require a suitable smoothing operation, using, say, a kernel smoother on $\hat{H}_0(\cdot)$. But this is not really required here, because our aim is to compare MSEs, and the very same multiplicative factor $h_0(\xi_0)^{-1}$ enters each of the Λ_S . Therefore, we may compare and estimate the MSEs of the variables $\bar{\Lambda}_S = h_0(\xi_0) \Lambda_S$, which has

$$\begin{aligned} E \bar{\Lambda}_S^2 &= \text{var } W(\xi_0) + \mathbf{R}_0^t \mathbf{J}_{00}^{-1} \mathbf{R}_0 + \boldsymbol{\zeta}^t \boldsymbol{\Omega}_S \mathbf{K} \boldsymbol{\Omega}_S^t \boldsymbol{\zeta} \\ &\quad + \boldsymbol{\zeta}^t (\mathbf{I} - \boldsymbol{\Omega}_S) \boldsymbol{\eta} \boldsymbol{\eta}^t (\mathbf{I} - \boldsymbol{\Omega}_S)^t \boldsymbol{\zeta}. \end{aligned}$$

Here the first three terms combine to give the variance part, whereas the fourth term stems from the model bias. Moreover, the two first terms are not affected by the model choice S , and they represent the minimal possible variance achieved by using the narrow model, where $S = \emptyset$. We see that the structure of these limiting risk expressions is of precisely the same form as in (14), as found there through Lemma 3, with the only essential difference being that

$$\begin{aligned} \boldsymbol{\zeta} &= \mathbf{R}_1 - \mathbf{J}_{10} \mathbf{J}_{00}^{-1} \mathbf{R}_0 \\ &= \frac{\partial f}{\partial \boldsymbol{\gamma}} - \mathbf{J}_{10} \mathbf{J}_{00}^{-1} \frac{\partial f}{\partial \boldsymbol{\beta}} - \{\mathbf{J}_{10} \mathbf{J}_{00}^{-1} \mathbf{F}_0(\xi_0) - \mathbf{F}_1(\xi_0)\} \end{aligned}$$

of Lemma 4 replaces $\boldsymbol{\omega} - \boldsymbol{\kappa}$ of Lemma 3. Clearly, $\boldsymbol{\zeta}$ is very similar to $\boldsymbol{\omega} - \boldsymbol{\kappa}$, but, as explained earlier, Lemma 3 does not cover the type of estimands handled by Lemma 4, which required a separate treatment.

Going back to the quantiles of the conditional survival time, for an individual with covariate $(\mathbf{x}, \mathbf{z})$, we have $\xi(r) = H_0^{-1}(f(\boldsymbol{\beta}, \boldsymbol{\gamma}))$ with $f(\boldsymbol{\beta}, \boldsymbol{\gamma}) = c \exp(-\mathbf{x}^t \boldsymbol{\beta} - \mathbf{z}^t \boldsymbol{\gamma})$ and $c = -\log(1-r)$ for the r th quantile. Thus $\boldsymbol{\zeta} = \mathbf{R}_1 - \mathbf{J}_{10} \mathbf{J}_{00}^{-1} \mathbf{R}_0$ with $\mathbf{R}_0 = \mathbf{F}_0(\xi_0) - c \exp(-\mathbf{x}^t \boldsymbol{\beta}) \mathbf{x}$ and $\mathbf{R}_1 = \mathbf{F}_1(\xi_0) - c \exp(-\mathbf{x}^t \boldsymbol{\beta}) \mathbf{z}$, and $c = \log 2$ for the case of the median.

6.3 Estimation of Risk Quantities

Theory and calculations developed in the previous sections led to the definition of various model-based quantities that need to be estimated in practice. This is required particularly in light of the model selection criteria of the next section. Here we describe the required estimators. We use $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ to denote the full-model-based partial likelihood estimators.

This is a convenient place to discuss estimation of τ_0 , $\mathbf{J} = \mathbf{J}_{\text{full}}$, $\mathbf{K}$, $\boldsymbol{\Omega}_S$, $\mathbf{K}_S$, $\boldsymbol{\omega}$, $\boldsymbol{\kappa}$, $\boldsymbol{\zeta}$. Use of the theory in the following sections demands that the estimators $\hat{\tau}_0$, $\hat{\mathbf{J}}$, and so on be consistent; that is, that they should converge in probability to the relevant quantities as n grows, under the local neighborhood circumstances (6). There are, in fact, several possibilities for say $\hat{\mathbf{J}}$ here, typically ranging from, say, $-\mathbf{I}_n(\hat{\boldsymbol{\beta}}_{\text{narr}}, \mathbf{0})$, which uses estimators from the narrow model, to $-\mathbf{I}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$, which uses estimators in the fullest $(p+q)$ -parameter model. The first-order large-sample theory that we develop does not distinguish between these estimators as long as they are consistent. In practice, we will typically prefer the full-model-based versions, partly for reasons of model robustness. (See CH and HC 2003 for a parallel discussion.)

We start with the information matrix $\mathbf{J}$, for which we use the full-model-based estimator (9). From this matrix we extract further estimates $\hat{\mathbf{J}}_{00}$, $\hat{\mathbf{J}}_{01}$, $\hat{\mathbf{J}}_{11}$, and $\hat{\mathbf{J}}_S$, and also $\hat{\mathbf{K}} = \hat{\mathbf{J}}^{11}$ with its consequent $\hat{\boldsymbol{\Omega}}_S = \pi_S^t \hat{\mathbf{K}}_S \pi_S \hat{\mathbf{K}}^{-1}$ matrices. We also use full-model-based parameter estimates for estimating the

$(p + q)$ -vector function $\mathbf{F}$ of Section 4, giving

$$\begin{aligned}\widehat{\mathbf{F}}(t) &= \int_0^t \mathbf{E}_n(u, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) d\widehat{H}_0(u) \\ &= \sum_{i=1}^n I\{t_i \leq t\} \mathbf{E}_n(t_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) \Delta \widehat{H}_0(t_i) \\ &= n^{-1} \sum_{i=1}^n I\{t_i \leq t\} \frac{\mathbf{G}_n^{(1)}(t_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})}{G_n^{(0)}(t_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})^2} \delta_i.\end{aligned}$$

Finally, there are a couple of options when estimating the partial derivatives of $\mu(\boldsymbol{\beta}, \boldsymbol{\gamma}, H_0)$, which are required to arrive at $\widehat{\boldsymbol{\omega}}$ and $\widehat{\boldsymbol{\kappa}}$. In most of our examples we are able to find explicit expressions for these, as for the examples given earlier in this section, after which we again insert parameter estimates from the fullest model. General numerical recipes also may be used in situations where explicit expressions are harder to come by.

For $\tau_0 = \tau_0(t)$ of (11), we insert estimates already described for the partial derivatives of μ with respect to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $H_0(t)$, and likewise for $\mathbf{F}_0(t)$ and $\mathbf{F}_1(t)$. The remaining integral $\int_0^t g^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})^{-1} dH_0(u)$ is estimated as

$$\begin{aligned}\int_0^t \frac{d\widehat{H}_0(u)}{G_n^{(0)}(u, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})} &= \sum_{i=1}^n I\{t_i \leq t\} \frac{\Delta \widehat{H}_0(t_i)}{G_n^{(0)}(t_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})} \\ &= n^{-1} \sum_{i=1}^n I\{t_i \leq t\} \frac{\delta_i}{G_n^{(0)}(t_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})^2}.\end{aligned}$$

For handling the model selection problems associated with conditional median or quantile survival distributions, or more generally situations where Lemma 4 is applicable, we need estimates of ξ_0 , $\mathbf{R}_0$, $\mathbf{R}_1$, and $\boldsymbol{\zeta}$. We use $\widehat{\xi}_0 = \widehat{H}_0^{-1}(f(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}))$, so for the median case we use $\widehat{H}_0^{-1}(c \exp(-\mathbf{x}^t \widehat{\boldsymbol{\beta}} - \mathbf{z}^t \widehat{\boldsymbol{\gamma}}))$ with $c = \log 2$. Similarly, we use

$$\begin{aligned}\widehat{\mathbf{R}}_0 &= \widehat{\mathbf{F}}_0(\widehat{\xi}_0) - c \exp(-\mathbf{x}^t \widehat{\boldsymbol{\beta}} - \mathbf{z}^t \widehat{\boldsymbol{\gamma}}) \mathbf{x} \quad \text{and} \\ \widehat{\mathbf{R}}_1 &= \widehat{\mathbf{F}}_1(\widehat{\xi}_0) - c \exp(-\mathbf{x}^t \widehat{\boldsymbol{\beta}} - \mathbf{z}^t \widehat{\boldsymbol{\gamma}}) \mathbf{z},\end{aligned}$$

also leading to the crucial quantity $\widehat{\boldsymbol{\zeta}} = \widehat{\mathbf{R}}_1 - \widehat{\mathbf{J}}_{10} \widehat{\mathbf{J}}_{00}^{-1} \widehat{\mathbf{R}}_0$.

7. THE FOCUSED INFORMATION CRITERION FOR COX REGRESSION MODELS

In this section we use theory developed in earlier sections to derive an FIC (that specified in Sec. 1.3) for general use in Cox proportional hazards regression models. The theory partly parallels that developed in CH and HC (2003), but required more buildup work (viz. Lemmas 1–4), due to the semiparametric nature of the Cox model and the fact that the partial likelihood does not involve the baseline hazard.

7.1 The Focused Information Criterion

We have demonstrated in Sections 5 and 6 that for focus estimands $\mu(\boldsymbol{\beta}, \boldsymbol{\gamma}, H_0)$, with ensuing estimators $\widehat{\mu}_S = \mu(\widehat{\boldsymbol{\beta}}_S, \widehat{\boldsymbol{\gamma}}_S, \widehat{H}_{0,S})$, the limiting risk can be expressed as

$$\text{risk}(S) = \tau_0^2 + (\boldsymbol{\omega} - \boldsymbol{\kappa})^t \{(\mathbf{I} - \boldsymbol{\Omega}_S) \boldsymbol{\eta} \boldsymbol{\eta}^t (\mathbf{I} - \boldsymbol{\Omega}_S)^t + \boldsymbol{\Omega}_S \mathbf{K} \boldsymbol{\Omega}_S^t\} (\boldsymbol{\omega} - \boldsymbol{\kappa})$$

for the relevant τ_0 , $\boldsymbol{\omega}$, and $\boldsymbol{\kappa}$. When the full $(p + q)$ -parameter model is used, for example, $\boldsymbol{\Omega}_S = \mathbf{I}_q$ and the risk function is

$\tau_0^2 + (\boldsymbol{\omega} - \boldsymbol{\kappa})^t \mathbf{K} (\boldsymbol{\omega} - \boldsymbol{\kappa})$, constant in $\boldsymbol{\eta}$. The other extreme is to select the narrow p -parameter model, for which $\boldsymbol{\Omega}_S = \mathbf{0}$, leading to risk function $\tau_0^2 + \{(\boldsymbol{\omega} - \boldsymbol{\kappa})^t \boldsymbol{\eta}\}^2$.

In these risk expressions, quantities τ_0 , $\boldsymbol{\omega}$, $\boldsymbol{\kappa}$, $\boldsymbol{\Omega}_S$, and $\mathbf{K}$ may all be estimated consistently with ordinary $\sqrt{n}$ precision; see the recipes of Section 6.3. The only quantity that cannot be estimated consistently is $\boldsymbol{\eta} \boldsymbol{\eta}^t$. For $\boldsymbol{\eta}$, about the best we can do is $\widehat{\boldsymbol{\eta}} = \sqrt{n} \widehat{\boldsymbol{\gamma}}_{\text{full}}$, using the estimator of $\boldsymbol{\gamma}$ in the full Cox model with all $p + q$ covariates included. From Lemma 1,

$$\widehat{\boldsymbol{\eta}} = \sqrt{n} \widehat{\boldsymbol{\gamma}}_{\text{full}} \rightarrow_d \mathbf{D} \sim N_q(\boldsymbol{\eta}, \mathbf{K}). \quad (16)$$

For the $\boldsymbol{\eta} \boldsymbol{\eta}^t$ quantity, therefore, the natural estimator is $\widehat{\boldsymbol{\eta}} \widehat{\boldsymbol{\eta}}^t - \widehat{\mathbf{K}} = n \widehat{\boldsymbol{\gamma}} \widehat{\boldsymbol{\gamma}}^t - \widehat{\mathbf{K}}$, in that $\widehat{\boldsymbol{\eta}} \widehat{\boldsymbol{\eta}}^t \rightarrow_d \mathbf{D} \mathbf{D}^t$, a variable with mean $\boldsymbol{\eta} \boldsymbol{\eta}^t + \mathbf{K}$. Thus we have an asymptotically unbiased risk estimator,

$$\begin{aligned}\widehat{\text{risk}}(S) &= \widehat{\tau}_0^2 + (\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}})^t \{(\mathbf{I} - \widehat{\boldsymbol{\Omega}}_S) (\widehat{\boldsymbol{\eta}} \widehat{\boldsymbol{\eta}}^t - \widehat{\mathbf{K}}) (\mathbf{I} - \widehat{\boldsymbol{\Omega}}_S)^t \\ &\quad + \widehat{\boldsymbol{\Omega}}_S \widehat{\mathbf{K}} \widehat{\boldsymbol{\Omega}}_S^t\} (\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}}) \quad (17)\end{aligned}$$

for each candidate model S . The FIC consists of selecting the model with smallest estimated risk.

It is useful in practice to compute each of these risk numbers, because they have direct interpretation as estimates of sample size times MSE. We could also usefully display $\text{FIC}^*(S) = \{\widehat{\text{risk}}(S)/n\}^{1/2}$, because these are estimates of root MSE. Statisticians are used to interpreting standard errors (i.e., estimated standard deviations), the ubiquitous companions to estimates of model parameters. Here we suggest also supplying the FIC^* numbers, supplemented with $\widehat{\text{FIC}}$ scores defined in the next paragraph.

As long as emphasis is on model selection, we may simplify the foregoing algebra somewhat and give a crisper, but equivalent, version of the FIC. For this, we subtract the constant $\widehat{\tau}_0^2$, which does not affect model comparison, and further subtract out the quantity $(\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}})^t \widehat{\mathbf{K}} (\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}})$, which is also common to each risk estimate. These rearrangements lead to

$$\widehat{\text{FIC}}(S) = (\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}_S)^2 + 2(\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}})^t \boldsymbol{\pi}_S^t \widehat{\mathbf{K}}_S \boldsymbol{\pi}_S (\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}}), \quad (18)$$

in terms of estimates $\widehat{\boldsymbol{\psi}} = (\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}})^t \widehat{\boldsymbol{\eta}}$ and $\widehat{\boldsymbol{\psi}}_S = (\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}})^t \widehat{\boldsymbol{\Omega}}_S \widehat{\boldsymbol{\eta}}$ (see CH 2003, sec. 4). This better conveys the statistical balancing game between modeling bias (the first term) and estimation variability (the second term). The FIC sees to it that an optimal model is selected for the particular task at hand; different estimands $\mu(\boldsymbol{\beta}, \boldsymbol{\gamma}, H_0)$ correspond to different $\boldsymbol{\omega} - \boldsymbol{\kappa}$ and different $\boldsymbol{\psi}$.

We have reached formulas (17) and (18) in the framework of Section 5, using in particular Lemma 3, covering a broad variety of situations. For the case of quantile survival time estimators, we need to use Lemma 4 rather than Lemma 3, with a more complicated limit distribution. However, as argued after Lemma 4, the same structure emerges when we work with $\widehat{\Lambda}_S = h_0(\xi_0) \Lambda_S$, which means that the foregoing FIC formulas still work, with $\widehat{\boldsymbol{\zeta}}$ replacing $\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}}$.

Remark 3. The risk estimate (17) is the sum of a variance and a squared bias estimate. These terms can, with a little algebra in combination with (19), be expressed as

$$\begin{aligned}V_n(S) &= \widehat{\tau}_0^2 + (\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}})^t \widehat{\boldsymbol{\Omega}}_S \widehat{\mathbf{K}} \widehat{\boldsymbol{\Omega}}_S^t (\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}}) \\ &= \widehat{\tau}_0^2 + (\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}})^t \boldsymbol{\pi}_S^t \widehat{\mathbf{K}}_S \boldsymbol{\pi}_S (\widehat{\boldsymbol{\omega}} - \widehat{\boldsymbol{\kappa}})\end{aligned}$$

and

$$\begin{aligned}(B^2)_n(S) &= (\hat{\omega} - \hat{\kappa})^t (\mathbf{I} - \hat{\Omega}_S) (\hat{\eta} \hat{\eta}^t - \hat{\mathbf{K}}) (\mathbf{I} - \hat{\Omega}_S)^t (\hat{\omega} - \hat{\kappa}) \\ &= n \{ (\hat{\omega} - \hat{\kappa})^t (\mathbf{I} - \hat{\Omega}_S) \hat{\gamma} \}^2 \\ &\quad - (\hat{\omega} - \hat{\kappa})^t (\hat{\mathbf{K}} - \pi_S^t \hat{\mathbf{K}}_S \pi_S) (\hat{\omega} - \hat{\kappa}),\end{aligned}$$

also demonstrating that $(B^2)_n(S)$ typically will increase with n , with size essentially determined by $(\omega - \kappa)^t (\mathbf{I} - \Omega_S) \eta$. Nevertheless, it can happen that the event negligible bias:

$$n \{ (\hat{\omega} - \hat{\kappa})^t (\mathbf{I} - \hat{\Omega}_S) \hat{\gamma} \}^2 < (\hat{\omega} - \hat{\kappa})^t (\hat{\mathbf{K}} - \pi_S^t \hat{\mathbf{K}}_S \pi_S) (\hat{\omega} - \hat{\kappa}) \quad (19)$$

occurs, in which case we choose to redefine $(B^2)_n(S)$ as 0, to avoid estimating the squared bias with a negative number. Thus we redefine $\widehat{\text{risk}}(S) = V_n(S)$ and

$$\begin{aligned}\widehat{\text{FIC}}(S) &= V_n(S) + (B^2)_n(S) - \hat{\tau}_0^2 - (\hat{\omega} - \hat{\kappa})^t \hat{\mathbf{K}} (\hat{\omega} - \hat{\kappa}) \\ &= (\hat{\omega} - \hat{\kappa})^t (\pi_S^t \hat{\mathbf{K}}_S \pi_S - \hat{\mathbf{K}}) (\hat{\omega} - \hat{\kappa})\end{aligned}$$

for those instances where the negligible bias event (19) occurs.

7.2 Securing Good Average Performance

The FIC apparatus introduced earlier is at the outset tailor made for optimal model selection when considering a single parameter of interest. One would not infrequently encounter focus parameters that depend on a covariate value or a time point that one next would wish to study across portions of the covariate space or time scale. We now show that the FIC machinery also yields methods for dealing with such problems.

For concreteness of illustration, consider the estimand $\mu = \mu(\mathbf{z}) = \exp\{(\mathbf{x} - \mathbf{x}_0)^t \boldsymbol{\beta} + (\mathbf{z} - \mathbf{z}_0)^t\}$ discussed in Section 6.1, but now viewed as a function of $\mathbf{z}$ with levels $\mathbf{x}_0$ and $\mathbf{x}$ kept fixed. We wish to select a submodel S that provides optimal precision for estimates $\hat{\mu}_S(\mathbf{z})$ across many values of $\mathbf{z}$. From Lemma 3, $\sqrt{n}\{\hat{\mu}_S(\mathbf{z}) - \mu(\mathbf{z})\} \rightarrow_d \Lambda_S(\mathbf{z})$, say, of the form given there, with $\omega = \omega(\mathbf{z})$ as calculated in (15). From this, we infer that if Q_n is some distribution of covariates $\mathbf{z}$ tending to some limit Q , then, under mild conditions,

$$\begin{aligned}\rho_n(S) &= n \int \{\hat{\mu}_S(\mathbf{z}) - \mu(\mathbf{z})\}^2 dQ_n(\mathbf{z}) \\ &\rightarrow_d \rho(S) = \int \Lambda_S(\mathbf{z})^2 dQ(\mathbf{z}),\end{aligned}$$

say. Hence the average risk of using $\hat{\mu}_S(\mathbf{z})$, say $\text{risk}_n(S) = E\rho_n(S)$, converges to

$$\begin{aligned}\text{risk}(S) &= E\rho(S) \\ &= \int [\tau_0^2 + \omega(\mathbf{z})^t \{(\mathbf{I} - \Omega_S) \eta \eta^t (\mathbf{I} - \Omega_S)^t + \Omega_S \mathbf{K} \Omega_S^t\} \\ &\quad \times \omega(\mathbf{z})] dQ(\mathbf{z}) \\ &= \tau_0^2 + \text{Tr} \left[\{(\mathbf{I} - \Omega_S) \eta \eta^t (\mathbf{I} - \Omega_S)^t + \Omega_S \mathbf{K} \Omega_S^t\} \right. \\ &\quad \times \left. \int \omega(\mathbf{z}) \omega(\mathbf{z})^t Q(d\mathbf{z}) \right].\end{aligned}$$

This can be estimated as in the previous section, plugging in consistent estimators of τ_0 , Ω_S , and $\mathbf{K}$ along with the distributions Q_n for Q and $\hat{\eta} \hat{\eta}^t - \hat{\mathbf{K}}$ for $\eta \eta^t$. This leads to a list of $\widehat{\text{risk}}(S)$ numbers to be minimized over candidate models S .

In the present case, with (15) for $\omega(\mathbf{z})$, the crucial $\int \omega(\mathbf{z}) \times \omega(\mathbf{z})^t Q(d\mathbf{z})$ matrix becomes

$$\begin{aligned}\mathbf{J}_{10} \mathbf{J}_{00}^{-1} (\mathbf{x} - \mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0)^t \mathbf{J}_{00}^{-1} \mathbf{J}_{01} + \int (\mathbf{z} - \mathbf{z}_0) (\mathbf{z} - \mathbf{z}_0)^t Q(d\mathbf{z}) \\ - \mathbf{J}_{10} \mathbf{J}_{00}^{-1} (\mathbf{x} - \mathbf{x}_0) (\bar{\mathbf{z}} - \mathbf{z}_0)^t - (\bar{\mathbf{z}} - \mathbf{z}_0) (\mathbf{x} - \mathbf{x}_0)^t \mathbf{J}_{00}^{-1} \mathbf{J}_{01},\end{aligned}$$

where $\bar{\mathbf{z}} = \int \mathbf{z} Q(d\mathbf{z})$. We could, for example, take Q_n to be the empirical distribution of $\mathbf{z}_1, \dots, \mathbf{z}_n$ and take $\mathbf{z}_0$ to be the average of these, in which case the estimated $\int \omega \omega^t dQ$ matrix becomes $\hat{\mathbf{J}}_{10} \hat{\mathbf{J}}_{00}^{-1} (\mathbf{x} - \mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0)^t \hat{\mathbf{J}}_{00}^{-1} \hat{\mathbf{J}}_{01} + \mathbf{S}_n$, with $\mathbf{S}_n$ being the empirical covariance matrix of the $\mathbf{z}_i$'s.

The foregoing FIC-averaging scheme is illustrated in Section 9.2 for the Danish skin cancer survival data. Note that the foregoing reasoning is general in nature, and can be applied with appropriate variations to the task of finding the best subset S for best average estimation of the nine decile survival times $\text{Su}^{-1}(j/10 | \mathbf{x}, \mathbf{z})$, for example.

8. MODEL AVERAGE ESTIMATORS AND THEIR PERFORMANCE

In the previous section we developed the FIC, to be used for model selection purposes in connection with any focus parameter μ of interest. It is also important to understand the statistical behavior of the resulting estimator-post-selection strategy. Such estimators take the form $\hat{\mu} = \hat{\mu}(\hat{S})$, say, where $\hat{S}$ is the randomly selected submodel. This is a special case of a more general class termed *compromise estimators* in HC (2003a). In this section we develop theory for such model averaging strategies for the Cox model. We also study the performance of AIC and BIC selectors and their associated smoothed-AIC and smoothed-BIC variants.

8.1 Model Averaging Estimators

When several candidate models are being considered, as earlier, a natural idea is to form compromise estimators that weight across models in a suitable fashion. Specifically, now consider

$$\hat{\mu} = \sum_S w_n(S | \hat{\eta}) \hat{\mu}_S, \quad (20)$$

where the weights depend on $\hat{\eta} = \sqrt{n} \hat{\gamma}_{\text{full}}$ of (16) and sum to 1. Any estimation-post-selection estimators that act on $\hat{\eta}$ to determine the final model are of this form, with weight 1 for the chosen submodel and weight 0 for the others. We show later that the AIC and FIC strategies are essentially of this form. The following “master theorem” may now be stated; the proof is given in the Appendix. The quantities τ_0 , ω , κ , Ω_S , and $\mathbf{K}$ are as defined at the start of Section 5.

Theorem 1. Suppose that the regularity conditions used in Lemmas 1, 2, and 3 continue to hold, and assume that the random weights $w_n(S | \hat{\eta})$ used in the compromise estimator (20) are such that the vector of $w_n(S | \hat{\eta})$ tends in distribution to the vector of $w(S | \mathbf{D})$, in terms of the limit $\mathbf{D}$ of $\hat{\eta}$ as in (16), where each $w(S | \mathbf{D})$ has at most a finite number of discontinuities in $\mathbf{D}$. Then

$$\sqrt{n}(\hat{\mu} - \mu_{\text{true}}) \rightarrow_d \Lambda = \Lambda_0 + (\omega - \kappa)^t \{\eta - \hat{\eta}(\mathbf{D})\}.$$

Here $\Lambda_0 \sim N(0, \tau_0^2)$ is independent of $\mathbf{D} \sim N_q(\eta, \mathbf{K})$, and $\hat{\eta}(\mathbf{D}) = \sum_S w(S | \mathbf{D}) \Omega_S \mathbf{D}$.

As a consequence of this result, for any model averaging estimator, we may compute its limit risk function under squared error loss as $\text{risk}(\eta) = E\Lambda^2 = \tau_0^2 + R(\eta)$, say, where

$$R(\eta) = E[(\omega - \kappa)^t \{\hat{\eta}(\mathbf{D}) - \eta\}]^2 \\ = (\omega - \kappa)^t E\{\hat{\eta}(\mathbf{D}) - \eta\} \{\hat{\eta}(\mathbf{D}) - \eta\}^t (\omega - \kappa). \quad (21)$$

Note the broad generality here; the performance of almost every model averaging strategy can be precisely assessed, for large n , by evaluating the precision of the estimator $(\omega - \kappa)^t \hat{\eta}(\mathbf{D})$ for the estimand $(\omega - \kappa)^t \eta$, in the limit experiment where $\mathbf{D} \sim N_q(\eta, \mathbf{K})$ and where all quantities are known apart from η . In particular, performance of the AIC versus that of the FIC, and so on, can be studied for different situations determined by $\mathbf{K}$ and $\omega - \kappa$.

Some natural model averaging estimators take the form $\sum_S w'_n(S) \hat{\mu}_S$, with data-dependent weights $w'_n(S)$ that are not direct functions of $\hat{\eta}$, as in (20). It follows from the proof of the theorem that the limit distribution is as described there as long as $w'_n(S) \rightarrow_d w(S | \mathbf{D})$ for each S . This remark particularly pertains to AIC and FIC strategies.

8.2 Akaike Information Criterion Performance for Cox Regression Models

The AIC is generally applicable for comparing competing parametric models. (See, e.g., Burnham and Anderson 2002 for a broad introduction with applications to many kinds of models.) The arguments behind its construction do not necessarily apply to the Cox regression model, however, because of its semiparametric nature. We are not aware of any other attempts in the literature to define or discuss aspects or performance of the AIC for the Cox model. We are, however, free to define and analyze

$$\text{AIC}_{n,S} = 2 \log L_{n,S}(\hat{\beta}_S, \hat{\gamma}_S) - 2(p + |S|) \quad (22)$$

in the style of parametric models, where $L_{n,S}$ is the partial likelihood function engaging (β, γ_S) [see (3)]. The submodel S with largest AIC score (22) is selected.

A useful representation of $\text{AIC}_{n,S}$ can be derived, in terms of $\hat{\eta}$ of (16), as

$$\text{AIC}_{n,S} - \text{AIC}_{n,\emptyset} = \hat{\eta}^t \mathbf{K}^{-1} \pi_S^t \mathbf{K}_S \pi_S \mathbf{K}^{-1} \hat{\eta} - 2|S| + o_p(1).$$

This may be shown following arguments in HC (2003a). Note in particular that all $\text{AIC}_{n,S}$ numbers across submodels S depend essentially only on the $\hat{\eta}$ vector. We may also show from this that

$$\text{AIC}_{n,S} - \text{AIC}_{n,\emptyset} \rightarrow_d \text{aic}(S) = \mathbf{D}^t \mathbf{K}^{-1} \pi_S^t \mathbf{K}_S \pi_S \mathbf{K}^{-1} \mathbf{D} - 2|S| \\ \sim \chi_{|S|}^2(\lambda_S) - 2|S|, \quad (23)$$

with eccentricity parameters $\lambda_S = \eta^t \mathbf{K}^{-1/2} \pi_S^t \mathbf{K}_S \pi_S \mathbf{K}^{-1/2} \eta$ and where $\mathbf{D}$ is as in (16). If $\mathbf{K}$ is diagonal $(k_1, \dots, k_q)$, then $\lambda_S = \sum_{j \in S} \eta_j^2 / k_j$.

These results imply in particular that there are well-defined precise limit probabilities for the different submodels being chosen by the AIC, specifically,

$$\text{ch}_n(S, \eta) = \Pr\{\text{model } S \text{ is chosen}\} \\ \rightarrow \Pr\{\text{aic}(S) \text{ is bigger than all other aic}(S')\} \\ = \text{ch}(S, \eta).$$

As an illustration, assume that we wish to select either the narrow model with only β or the fullest model with all of (β, γ) . Then

$$\text{ch}_n(\text{full}, \eta) \rightarrow \text{ch}(\text{full}, \eta) = \Pr\{\chi_q^2(\eta^t \mathbf{K}^{-1} \eta) \geq 2q\},$$

a probability that increases with the distance between η and $\mathbf{0}$. In view of (6), the probability of selecting the fullest model will typically increase toward 1 with sample size n , with a speed dictated by the size of $\gamma^t \mathbf{K}^{-1} \gamma$.

It is worth pointing out that the AIC as developed here, using the partial likelihood, in a sense does not consider H_0 or how well the indirectly selected $\hat{H}_{0,S}$ estimator performs. The FIC methods of Section 7, on the other hand, may be geared toward good estimation performance for H_0 , for example.

Theorem 1 gives the limit distribution of AIC estimators, corresponding to the particular random weights $w_a(S | \mathbf{D}) = I\{\text{aic}(S) \text{ is larger than all other aic}(S')\}$. For the case of $\mathbf{K}$ diagonal, the winning subset S consists of precisely those j for which $D_j^2 / k_j > 2$, which means that the AIC essentially acts as a pretest procedure with significance level $\Pr\{\chi_1^2 > 2\} = .157$ per parameter tested. The theorem similarly yields the limit distribution of each estimator post-FIC, with $w_f(S | \mathbf{D})$ equal to 0 for all sets S except the one in which

$$\text{fic}(S) = (\omega - \kappa)^t \{(\mathbf{I} - \Omega_S)(\mathbf{D}\mathbf{D}^t - \mathbf{K})(\mathbf{I} - \Omega_S)^t + \Omega_S \mathbf{K} \Omega_S^t\} \\ \times (\omega - \kappa)$$

is the smallest. We may now compare AIC and FIC performance in given settings, through (21) and the respective $\hat{\eta}(\mathbf{D}) = \sum_S w(S | \mathbf{D}) \Omega_S \mathbf{D}$ estimators of η .

8.3 Smoothing the Akaike Information Criterion, Bayes Information Criterion, and Focused Information Criterion

Theorem 1 is very general and allows a broad class of model averaging estimators. Among these, we single out two procedures: smoothed versions of the AIC and the FIC. Further options were discussed in HC (2003a). For the smoothed AIC, we use (20) with data-determined weights,

$$w_{n,\text{aic}}(S) = \frac{\exp(\frac{1}{2}\lambda \text{AIC}_{n,S})}{\sum_{\text{all } S'} \exp(\frac{1}{2}\lambda \text{AIC}_{n,S'})}, \quad (24)$$

in terms of the AIC scores (22). The sum in the denominator extends over all submodels under consideration. The list of these does not have to be extensive; often, some submodels can be ruled out on a priori grounds. The λ of (24) is like a smoothing parameter, dictating the amount of smoothing between candidate models. If λ is large, then the method is essentially equivalent to the AIC selection scheme; if $\lambda = 0$, then all methods are weighted equally. Certain arguments provided by Buckland, Burnham, and Augustin (1997) and Burnham and Anderson (2002) on an ad hoc basis and more fully in HC (2003a) advocate taking $\lambda = 1$ in (24). This is the value that we take in our simulations and illustrations in Section 9. The limit distribution of the smoothed AIC estimator is as in Theorem 1 with $w(S | \mathbf{D})$ proportional to $\exp\{\frac{1}{2}\lambda \text{aic}(S | \mathbf{D})\}$, in view of (23) and the remark ending Section 8.1.

In a similar manner, the smoothed FIC uses (20) with weights

$$w_{n,\text{fic}}(S) = \exp \left\{ -\frac{1}{2} \lambda \frac{\widehat{\text{FIC}}(S)}{(\widehat{\omega} - \widehat{\kappa})^t \widehat{\mathbf{K}}(\widehat{\omega} - \widehat{\kappa})} \right\} / \sum_{\text{all } S'} \exp \left\{ -\frac{1}{2} \lambda \frac{\widehat{\text{FIC}}(S')}{(\widehat{\omega} - \widehat{\kappa})^t \widehat{\mathbf{K}}(\widehat{\omega} - \widehat{\kappa})} \right\}, \quad (25)$$

again with a parameter λ determining the degree to which lower FIC scores should be compared with higher ones. The point of the scaling here, through the $(\widehat{\omega} - \widehat{\kappa})^t \widehat{\mathbf{K}}(\widehat{\omega} - \widehat{\kappa})$ factor, is to make different situations similar with respect to the scale of the smoothing parameter λ ; $(\omega - \kappa)^t \mathbf{K}(\omega - \kappa)$ is the constant risk of the minimax method in the limit experiment alluded to after Theorem 1. In our illustrations, we have taken $\lambda = 1$. Larger values would push the model averaging method closer to the FIC method, and values closer to 0 would correspond to equal weights across the submodels considered. The limit distribution is again given by the theorem, with weights $w(S | \mathbf{D})$ proportional to $\exp\{-\frac{1}{2} \lambda' \text{fic}(S)\}$, where $\lambda' = \lambda / \{(\omega - \kappa)^t \mathbf{K}(\omega - \kappa)\}$.

Variations exist, such as using (24) and (25) involving, say, only the top 10 marked models. Theorem 1 still applies and accurately describes the large-sample performance of such model averaging schemes. Limit distributions are nonlinear mixtures of normals and as such are nonnormal; see the illustration in Section 10.6.

As with (24) and (25), we may form smoothed BIC estimators by simply replacing the AIC scores in (24) with those of the BIC. Smoothing BIC scores is actually somewhat different from smoothing AIC and FIC scores. Inside our local neighborhood framework [cf. (6)], arguments that led to (23) imply that $\text{BIC}_{n,S} - \text{BIC}_{n,\emptyset} = Z_n(S) - |S| \log n + o_p(1)$, where $Z_n(S)$ converges in distribution. It follows that BIC penalizes sufficiently heavily for complexity when n grows for the narrow model $S = \emptyset$ to be the winner, with probability converging to 1. (See Claeskens and Hjort 2004 for some more formal arguments and further discussion in a different but similar context.) In other words, $w_{n,\text{bic}}(\emptyset) \rightarrow_p 1$, and the smoothed BIC estimator becomes asymptotically equivalent to the narrow model strategy. The limit risk is $\text{risk}(\eta) = \tau_0^2 + \{(\omega - \kappa)^t \eta\}^2$, by (21), and in particular is unbounded in η .

8.4 Consistency May Be a Bad Property

Different issues may be addressed when studying performance aspects of model selectors. One issue is that of having high probability for selecting the “right” model. Some methods have the consistency property, that is, the probability of selecting the right model goes to 1 when the sample size increases. The BIC of (2), as well as the SCAD method of Fan and Li (2002) and the penalized likelihood method of Bunea and McKeague (2005), have this “oracle” property. If a certain submodel is correct, then the model selector will find this submodel with probability going to 1, and the final estimator after selection will be asymptotically equivalent to the estimator constructed under this submodel.

In this article we are more concerned with another issue: performance of the final estimator. Some inspection reveals that the oracle property sounds better than what it actually delivers, and that there are unfortunate side effects regarding estimator performance. We discuss a certain max-risk problem associated

with the BIC, the SCAD, and the penalized likelihood method in the present context of proportional hazards regression models, a problem to which the FIC and FIC-smoothed methods that we have developed are immune. This disturbing max-risk phenomenon has been discussed by Foster and George (1994) for the BIC in multiple regression, by Yang (2005) for more general oracle procedures in a context of additive regression models, and by Leeb and Pötscher (2004, 2005). The latter authors worked inside a somewhat limited framework of subset selection in linear regression models (with a fixed and known value for the residual standard deviation parameter σ , in some conflict with the aspect of modeling that larger models entail smaller σ and vice versa), but argued that the phenomena exhibited there should occur rather more generally.

To describe what happens in the Cox regression framework it will suffice to consider the case of $p = 0$ and $q = 1$, so there are only two models, M_0 and M_1 , to consider, corresponding to $\gamma = 0$ and $\gamma \in \mathcal{R}$. Several model selectors are of the form

$$\text{select} \quad \begin{cases} M_0 & \text{if } |V_n| \leq c_n^{1/2} \\ M_1 & \text{if } |V_n| > c_n^{1/2} \end{cases}$$

or are asymptotically equivalent to such a form, where $V_n = \sqrt{n} \widehat{\gamma} / \widehat{\sigma}$ in terms of the full-model Cox estimator $\widehat{\gamma}$ and its estimated standard deviation $\widehat{\sigma} / \sqrt{n}$. That both the AIC and BIC methods can be paraphrased as here for $c_n = 2$ and $c_n = \log n$ follows from elaborations of the (2) formulas in the style of results of Section 8.2. Moreover, each FIC method is essentially equivalent to the AIC here, when $q = 1$ (see CH 2003, sec. 4). Consistency outside the null model means that

$$P_n(M_1 | \gamma) = \Pr\{|V_n| > c_n^{1/2} | \gamma\} \rightarrow 1 \quad \text{for } \gamma \neq 0,$$

which is seen to be equivalent to $c_n/n \rightarrow 0$, so both AIC and BIC have this property. But consistency at the null model takes $P_n(M_1 | \gamma) \rightarrow 0$ for $\gamma = 0$, which is seen to mean $c_n \rightarrow \infty$. So BIC is null-model consistent, whereas the AIC and FIC are not.

That consistency may be unfortunate is seen by examining the associated estimator post selection,

$$\mu^* = \begin{cases} \mu(0) & \text{if } |V_n| \leq c_n^{1/2} \\ \mu(\widehat{\gamma}) & \text{if } |V_n| > c_n^{1/2}. \end{cases}$$

Take, for example, $\mu = \exp(\gamma)$, the relative risk of individuals with covariate $z + 1$ compared with those with covariate z . Then

$$\begin{aligned} R_n(\gamma) &= nE(\mu^* - \mu)^2 \\ &= nE[\exp(\widehat{\gamma}) I\{|V_n| > c_n^{1/2}\} \\ &\quad + I\{|V_n| \leq c_n^{1/2}\} - \exp(\gamma)]^2. \end{aligned}$$

We may show here that this risk function is uniformly bounded in n and γ , provided that c_n stays bounded (as for the AIC and FIC), whereas it becomes highly unstable for γ close to 0 when $c_n \rightarrow \infty$ (as for the BIC). Some analysis shows that $R_n(a(c_n/n)^{1/2}) \rightarrow \infty$, for $a = \frac{1}{2}\sigma$, where σ is the limit in probability of $\widehat{\sigma}$; the key is that then $|V_n| \leq c_n^{1/2}$ with probability converging to 1. It may be similarly demonstrated that the max-risk for any oracle property method goes to infinity with n , whereas the max-risk of the AIC and of any FIC method stays bounded.

9. ILLUSTRATIONS AND APPLICATIONS

In this section we illustrate our FIC and model averaging methods in simulations and find that the postselection FIC as well as the smoothed FIC methods may perform well in comparison with, for example, the AIC and BIC regimes. We then analyze the Danish skin cancer survival study data described in Section 1.1.

9.1 Results of a Simulation Study

Data t_i are generated following a Cox proportional hazards regression model,

$$h_i(u) = h_0(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\gamma}) = h_0(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\eta} / \sqrt{n}),$$

for $i = 1, \dots, n$, with constant baseline hazard $h_0(t) = 1$ and with all $p + q$ covariates $x_{i,j}$ and $z_{i,j}$ generated independently from the standard normal distribution. In each setting we use $p = 2$ protected variables with $\boldsymbol{\beta} = (1, 1)^t$. Censoring times are generated from an exponential distribution with mean 10/9. In our settings this corresponds to an average proportion of censored observations of about 52%. This figure is approximately valid for each of our settings, because $\boldsymbol{\gamma} = \boldsymbol{\eta} / \sqrt{n}$ is not far from 0. In setting (A), $q = 4$, and data are generated under the narrow model assumption, that is, $\boldsymbol{\eta} = (0, 0, 0, 0)^t$. Situation (B) is the same as (A) but corresponding to the full model with $\boldsymbol{\eta} = (3, -3, 3, -3)^t$. In setting (C) a situation in between the narrow and the full model is taken with $q = 6$ but $\boldsymbol{\eta} = (0, 0, 3, -3, 3, -3)^t$.

In these situations we study four focus parameters. Focus parameter (a) is the relative risk of a subject with covariates at value .5, relative to the mean covariate values, i.e. $\mu(\mathbf{x}, \mathbf{z}) = \exp(\mathbf{x}^t \boldsymbol{\beta} + \mathbf{z}^t \boldsymbol{\gamma})$ with each $\mathbf{x}$ and each $\mathbf{z}$ fixed at .5. Focus point (b) is $H_0(t)$, the cumulative baseline hazard rate function at time $t = .5$. Our third focus (c) is the survival probability $\text{Su}(t | \mathbf{x}, \mathbf{z})$ for a subject with the same covariates and time values as in (a) and (b). The last focus point (d) is the median survival time $\xi(.5 | \mathbf{x}, \mathbf{z})$, with again the same covariate values as in (a).

For each of the $\text{sim} = 10,000$ simulation runs, we compute for all subsets the estimators of the focus parameters μ in each of the 2^q models, together with the values of AIC, BIC [per (2)], and for each focus parameter the corresponding FIC. The final estimators are the post-model selection estimators pAIC, pBIC, and pFIC, as well as the model-averaged weighted estimators wAIC, wBIC, and wFIC, with model weights based on the values of the information criteria for that model; see Section 8.2. We compute the root MSEs, $\{\text{sim}^{-1} \sum_{j=1}^{\text{sim}} (\hat{\mu}_j - \mu_{\text{true}})^2\}^{1/2}$, across the simulations. Two sample sizes are used, $n = 150$ and $n = 300$. Thus the true values of the four parameters are (2.718, .500, .257, .255) for setting (A) for both sample sizes and (3.200, .500, .202, .217) and (3.051, .500, .218, .227) for settings (B) and (C) for the two sample sizes. The results are summarized in Table 1. The winning criterion in each situation is identified with its score given in bold. Note that even with $\text{sim} = 10,000$ runs, there is some simulation uncertainty that leaves some of the comparisons still open and the identified “winners” not quite clear.

In setting (A), the narrow model is the true model. It is known that the BIC is a consistent model selector and that it

Table 1. Root MSEs Over 10,000 Simulation Runs [for (A) and (B)] and 3,000 Runs [for (C)] of the Post-Model Selection and Model-Averaged Estimators Based on the FIC, AIC, and BIC, for Focus Parameters (a) Relative Risk, (b) Cumulative Hazard, (c) Survival Probability, and (d) Median

		wFIC	pFIC	wAIC	pAIC	wBIC	pBIC
Setting (A)							
$n = 150$	(a)	.447	.456	.455	.489	.396	.401
	(b)	.090	.090	.090	.090	.089	.089
	(c)	.062	.063	.063	.066	.059	.060
	(d)	.048	.043	.049	.051	.045	.046
$n = 300$	(a)	.277	.277	.281	.301	.246	.249
	(b)	.063	.063	.063	.063	.063	.063
	(c)	.043	.043	.044	.046	.041	.041
	(d)	.033	.030	.033	.035	.031	.032
Setting (B)							
$n = 150$	(a)	.504	.535	.599	.638	.593	.692
	(b)	.091	.092	.093	.094	.092	.093
	(c)	.070	.076	.078	.081	.078	.088
	(d)	.063	.057	.061	.064	.064	.073
$n = 300$	(a)	.306	.319	.370	.393	.372	.434
	(b)	.063	.063	.064	.064	.063	.064
	(c)	.047	.051	.053	.055	.054	.060
	(d)	.039	.036	.041	.042	.042	.048
Setting (C)							
$n = 150$	(a)	.566	.582	.669	.731	.627	.722
	(b)	.091	.092	.093	.094	.092	.094
	(c)	.074	.078	.082	.087	.081	.090
	(d)	.069	.062	.065	.068	.068	.076
$n = 300$	(a)	.333	.330	.398	.422	.388	.450
	(b)	.063	.064	.064	.064	.064	.064
	(c)	.048	.050	.053	.056	.053	.059
	(d)	.041	.037	.040	.042	.041	.047

NOTE: Setting (A) corresponds to $\boldsymbol{\eta} = (0, 0, 0, 0)^t$; setting (B) to $\boldsymbol{\eta} = (3, -3, 3, -3)^t$; and setting (C) to $\boldsymbol{\eta} = (0, 0, 3, -3, 3, -3)^t$.

often works well for models with a small number of parameters. Hence it is expected to do well for this situation, as indeed it does. Note that especially for focus parameters (b) and (d), the differences with the FIC values are only minor. For settings (B) and (C), where there are four more nonzero parameters in the true model than in the narrow model, the BIC is no longer preferred. The smoothed FIC is clearly the best choice for setting (B) where the wide model is true, whereas for the median as a focus point, the post-FIC selector gives the best results. For setting (C), in between the narrow and full model, the pFIC and wFIC outperform the AIC and BIC. Overall, model averaging tends to yield smaller simulated MSE than post-model selection. Considering only the post-model selection estimators, we see that the pBIC performs well for the simplest setting where all extra parameters are 0. For setting (B) the pFIC is the best. pFIC is also the best for setting (C), where all three criteria perform about equally well for focus (b).

9.2 Survival Analysis for Malignant Melanoma

Here we examine the dataset described in Section 1, involving $n = 205$ patients with covariates $x_1, z_1, \dots, z_6$. As explained earlier, x_1 is a protected covariate, included in each candidate model, whereas all $2^6 = 64$ subsets of the z_j 's are searched through for our model selection tasks. The seven hazard regression coefficient estimates were .535 (.277) for β_1 , .036 (.052) for γ_1 , .321 (.192) for γ_2 , -.707 (.314) for γ_3 , -.995 (.324) for γ_4 , .334 (.241) for γ_5 , and .017 (.008) for γ_6 , with the esti-

mated standard deviation (standard error) in parentheses, computed using the full model. In particular, variables x_1 , z_2 , z_3 , z_4 , and z_6 may be considered to have a reasonably clear influence on lifetimes, as measured by the ratio estimates divided by standard error. Coefficients γ_1 and γ_5 would not be seen as significantly different from 0 in most analyses, however. Nevertheless, we see that variable z_5 will often be selected by different criteria for different purposes.

The model selection methods applied are the AIC, the BIC, and four versions of FIC, each corresponding to a particular focus parameter. FIC1 corresponds to $\mu_1 = \exp\{(\mathbf{x} - \mathbf{x}_0)^t \boldsymbol{\beta} + (\mathbf{z} - \mathbf{z}_0)^t \boldsymbol{\gamma}\}$, relative risk of a man with average tumor thickness among all men participating in the study, infection infiltration level $z_2 = 3$, epithelioid cells not present ($z_3 = 2$), ulceration present ($z_4 = 1$), invasion depth $z_5 = 1$, and average men's age in the study, compared with that of women with averages of tumor thickness and age computed over the subgroup of women in the study, with the other covariates the same. The first set of covariates for men defines the variables level $(\mathbf{x}, \mathbf{z})$, and the second set corresponds to $(\mathbf{x}_0, \mathbf{z}_0)$. FIC2 computes the FIC values for $\mu_2 = H_0(t)$ at time $t = 3 \cdot 365$ days, which corresponds to the time at which the estimated Kaplan-Meier survival probability reaches .85. The third focus, which defines FIC3, is the survival probability at time $t = 3 \cdot 365$ for the same set of covariates $(\mathbf{x}, \mathbf{z})$ as for the first focus parameter, that is, $\mu_3 = \text{Su}(t | \mathbf{x}, \mathbf{z})$. The final focus parameter is $\mu_4 = \xi(.10) = \text{Su}^{-1}(.90 | \mathbf{x}, \mathbf{z})$, the time at which at least 90% of the patients with covariate level $(\mathbf{x}, \mathbf{z})$ are still alive.

Table 2 shows the 10 highest-ranked score values for each of the criteria in question, after having searched through all $2^6 = 64$ candidate models, with the selected z_j variables. Here, for example, "245" means that variables z_2 , z_4 , and z_5 are included, and so on. Note that the values are sorted in terms of importance per criterion.

The AIC model choice method yields a model with the five variables z_2 , z_3 , z_4 , z_5 , and z_6 ; only z_1 , tumor thickness, is not selected. The BIC, on the other hand, selects only variables z_4 (ulceration) and z_5 (invasion depth). Variable z_4 is present in all of the 20 best BIC models. With the FIC, there is not one single answer for the "best" model, as explained in Sections 1 and 7; the model chosen by FIC depends on the focus. The relative risk as a focus parameter lets FIC point to the narrow model, followed by models with variables z_2 and then $\{z_2, z_6\}$, and so on. In the second example, to estimate the cumulative hazard $H_0(t)$ at time t equal to 3 years, only variable z_5 (invasion

Table 3. For Focus Parameter (3) and Survival Probability $\text{Su}(t | \mathbf{x}, \mathbf{z})$, for Each of the 10 Best Models According to FIC, as Well as for AIC (one but last line) and BIC (last line), the Variables in the Model, the Estimate $\hat{\mu}_S$, Classical Standard Error Pretending This Is the True Model, Estimated Bias, Root MSE, and the Value of FIC

Variables	$\hat{\mu}_S$	SE_S	$Bias_S$	$(MSE_S/n)^{1/2}$	FIC_S
5	.487	.048	.022	.048	.146
6	.431	.050	.014	.050	.184
46	.588	.050	.019	.050	.184
16	.523	.050	.020	.050	.185
56	.264	.051	-.004	.051	.189
146	.608	.050	.022	.051	.190
456	.425	.051	.005	.051	.190
156	.387	.051	.007	.051	.195
1456	.493	.051	.012	.051	.195
45	.640	.048	.030	.051	.200
23456	.286	.055	-.004	.055	.275
45	.640	.048	.030	.051	.200

depth) is selected, with the second-best model being $\{z_2, z_5\}$. The picture is somewhat different when studying the survival probability, focus μ_3 . Here variable z_5 is again the most important, followed by the model $\{z_4, z_5\}$, and so on. Note that variable z_6 (age) shows up quite frequently in the list of the best FIC3 models, whereas variable z_5 (invasion depth) appears to be the most important one for FIC2. For the 10% quantile [i.e., $\text{Su}^{-1}(.90 | \mathbf{x}, \mathbf{z})$], we find the narrow model, including none of the extra variables, to be the best FIC4 choice, with models $\{z_5\}$ and $\{z_5, z_6\}$ as second best.

The fact that different models are selected for different purposes should not lead to confusion; rather, it should be viewed as a way of strengthening the biostatistician's ability to produce more precise estimates or predictions for a specific patient or patient group.

Table 3 gives results for the model selection problem for focus parameter (3), the survival probability for a patient with covariates $(\mathbf{x}, \mathbf{z})$. We observe in particular that for the 10 best models, according to FIC3, the estimates of the focus parameter $\hat{\mu}_S = \text{Su}(t | \mathbf{x}, \mathbf{z})$ are quite variable, ranging from .209 (variables 5, 6) to .590 (variables 4 and 5). For other situations the estimates for the best say 10 models may be more homogeneous than for this particular case. Note that one version of the smoothed FIC method described in Section 8.2 is to take a weighted average of the 10 best candidate estimates $\hat{\mu}_S$ of Table 3, with weights as given there.

Whereas the arguably most important output from such an FIC analysis would be this list of the most important models,

Table 2. Values of the AIC, BIC, and FIC for Four Focus Parameters: (1) Relative Risk, (2) Cumulative Hazard, (3) Survival Probability, and (4) 10% Quantile $\xi(.10) = \text{Su}^{-1}(.90)$

Variables	AIC	Variables	BIC	Variables	FIC1	Variables	FIC2	Variables	FIC3	Variables	FIC4
23456	-527.15	45	-542.72	$\emptyset$	2.84	5	3.33	5	.15	$\emptyset$	.029
3456	-528.29	14	-542.98	2	4.11	25	4.09	6	.18	5	.038
12346	-528.59	345	-544.04	26	4.17	235	4.11	46	.18	56	.038
123456	-528.69	24	-544.71	25	4.30	56	4.15	16	.18	15	.039
2346	-529.57	4	-544.89	256	4.35	356	4.15	56	.19	25	.039
13456	-529.65	3456	-544.91	6	4.63	125	4.40	146	.19	156	.040
2345	-529.94	124	-545.05	456	4.85	35	4.42	456	.19	256	.040
1234	-530.45	134	-545.42	14	5.22	2	4.43	156	.20	3	.042
1346	-530.51	456	-545.47	145	5.22	156	4.46	1456	.20	23	.042
345	-530.74	245	-545.49	146	5.26	1356	4.46	45	.20	13	.042

NOTE: The table shows the 10 largest AIC and BIC values and the 10 smallest FIC values for each of the focus parameters. The values are sorted for each criterion and for each focus parameter; the selected variables among $z_1, \dots, z_6$ are listed.

along with FIC scores and the list of corresponding $\hat{\mu}$ estimates, it is also often fruitful to examine the bias and standard deviation components that combine to give the risk estimate (17). As demonstrated in Section 7, this risk estimate is directly related to the FIC formula (18), and sorting models by the FIC score is equivalent to sorting by risk estimate. The bias can be estimated in a couple of different ways. What we can term the “direct bias estimate” involves plugging in estimates in the bias formula in Lemma 3 of Section 5; that is, using bias_S equal to $(\hat{\omega} - \hat{\kappa})^t(\mathbf{I} - \hat{\Omega}_S)\hat{\eta}/\sqrt{n}$. We divide by $\sqrt{n}$ here to give an estimate for the genuine bias of $\hat{\mu}_S$ at sample size n . The alternative method is as spelled out in Remark 3, with bias_S equal to the signed square root of $(B^2)_n(S)/\sqrt{n}$. It is with this second version, through the correct estimate of the squared bias, that we have the Pythagorean combination of bias_S and se_S giving $\text{mse}_S/\sqrt{n}$.

The FIC reflects the bias–variance trade-off phenomenon, as seen here by having larger biases for simpler models and larger standard errors for more complex models, even though the standard errors and resulting MSE estimates are fairly close in this particular situation.

We take the FIC analysis of this dataset one step further by examining the average FIC and risk criteria developed in Section 7.3 to select models with good performance across user- and context-defined portions of the covariate space. Specifically, we study focused model selection for the relative risk of a man with average tumor thickness among all persons participating in the study, infection infiltration level $z_2 = 3$, epithelioid cells not present ($z_3 = 2$), ulceration present ($z_4 = 1$), invasion depth $z_5 = 1$, compared with the risk of women with covariate information (x_0, z_0) as given earlier in this section. This corresponds to a certain $\mu(x, z, x_0, z_0)$ parameter, with $x = 2$ for man versus $x = 1$ for women. The present task is to find the best submodel for the best average performance, weighted across all ages z_6 . Following the method of Section 7.3, we compute, for all 205 subjects in the study ranging in age from 4 to 95, and for each of the $2^6 = 64$ models S , the values of FIC and the corresponding $(\text{MSE}_S/n)^{1/2}$. The resulting 205×64 FIC values can be analyzed by patient. Summarizing this, we have the following situation:

Variables:	1	24	26	2	35	3	46	4	5	$\emptyset$
Times selected:	10	37	3	7	7	12	12	35	10	72.

This corresponds to individual selection per patient or per age. Because there is clearly a difference in individual model choice, for an overall model we follow the approach of Section 7.3 and compute both the average FIC_S and the average $\text{risk}(S)$ values, taken over all 205 patients, then order the resulting 64 averaged values. Results for the five best models, given in Table 4, show that the overall model with variables z_2 and z_6 is the best choice. This application points yet again to the importance of first considering using the selected model before blindly applying a model selection criterion.

The picture does not always have to be so diverse. If we carry out the same exercise for focus parameters μ_2 , then the FIC2 best model is chosen in all 205 cases, and the model for best average performance coincides with the individual best models. This gives a good overall model for μ_2 , independent of age.

Table 4. For Focus Parameter (1), the Relative Risk as Specified in the Text, We Give for Each of the Five Best Models According to Averaged FIC the Variables in the Model, the Value of the Averaged FIC, and of the Averaged Risk

Variables	Averaged FIC1	$(\text{MSE}_S/n)^{1/2}$
26	24.92	30.079
456	25.44	30.088
346	26.23	30.102
126	26.37	30.104
1346	26.39	30.105

10. CONCLUDING REMARKS

We end the article with some comments and remarks, some of which might point to further research work.

10.1 “Protected” versus “Open”

Our framework uses p “protected” covariates, designated to always be inside the chosen models, with further q “open” covariates that may or may not be included in the final models. Deciding which is which is context-related and up to the statistician. For the analysis of Section 9.2 we could have chosen to use both sex x_1 and age z_6 as protected, for example, leaving the FIC and model average machinery to work with z_1, z_2, z_3, z_4 , and z_5 as open. The R software programs that we have developed make it easy to work through each desired combination. One may also choose not to pinpoint any protected covariates at all; that is, using $p = 0$ and leaving it all to data to decide which covariates should be included for what purposes. Our Lemmas 1–4 with Theorem 1 can be extended to cover the required $p = 0$ case.

10.2 Pretesting, Backward and Forward Selection

A simple pretesting approach is sometimes followed for covariate inclusion. For the skin cancer data one might include those z_j for which $|W_{n,j}| = |\hat{\gamma}_j/\text{se}(\hat{\gamma}_j)|$ exceeds, say, 1.645, corresponding to test significance level of .10 per coefficient; this yields the $\{z_2, z_3, z_4, z_6\}$ model, with z_1 and z_5 excluded. One might likewise have attempted versions of “forward” and “backward” selection strategies. It is important to realize that each of these methods is covered by Theorem 1 of Section 8, with appropriate nonnormal limit distributions. This also makes it possible to compare performances via the risk function $R(\eta)$ given there. A systematic study of this sort would be useful. A tentative and partial conclusion from some evidence presented in CH and HC (2003) is that the AIC and the FIC will tend to outperform the pretest regime, and that the smoothed versions wAIC and wFIC often are even better.

10.3 Two Cultures

There are at least two uses of statistical modeling of the Cox regression variety (and more general models). The models may be used primarily for interpretation (perhaps of a biomedical nature, e.g., finding that high blood pressure is associated with higher risk) and may be used primarily as a vehicle for producing precise estimation of quantities, such as median survival time, relative risk, survival probabilities, and so on. Though these uses are not fully orthogonal (see, e.g., Breiman 2001), the viewpoint of the present article has (primarily) been the

second one, aiming for as precise estimation and prediction as possible.

10.4 Even More Submodels

We have studied estimators of the form $\mu(\hat{H}_{0,S}, \hat{\beta}_S, \hat{\gamma}_S)$, but may in principle use different covariate subsets for the H_0 part and the (β, γ) part. Our techniques allow extensions to cover the required simultaneous distribution of all 2^{2q} combined estimators, which would lead to a somewhat more general FIC, and so on, but we do not pursue this here.

10.5 Bayesian Model Averaging

Section 8 dealt with the large class of compromise or model averaging estimators, with Theorem 1 being a “master theorem” giving the limit distribution of all such estimators. It can be further generalized in the direction of “generalized ridge” estimators (as in HC 2003a, secs. 8 and 9). These take the form

$$(\tilde{\beta}_S, \tilde{\gamma}_S) = (\hat{\beta}_S, \varepsilon(S | \hat{\eta}) \hat{\gamma}_S) \quad \text{with shrinking factor} \\ \varepsilon(S | \hat{\eta}) \in [0, 1],$$

with consequent $\tilde{\mu}_S = \mu(\tilde{\beta}_S, \tilde{\gamma}_S, \tilde{H}_{0,S})$. They may be particularly useful when the number q of extra covariates is growing large. It turns out that a class of Bayesian model averaging methods, as worked with from several perspectives by Volinsky, Madigan, Raftery, and Kronmal (1997), Clyde (1999), Hoeting, Madigan, Raftery, and Volinsky (1999), Volinsky and Raftery (2000), becomes first-order equivalent to a special subclass of such generalized ridge estimators. Thus their performance can be studied along with that of the frequentist model averaging schemes of Section 8.2.

10.6 Confidence Intervals and Tests

We have developed methods for selecting the “right” variables in applications with the Cox regression model, with different optimal subsets for different contexts and uses. Our methodology also extends to averages over candidate models and reaches precise descriptions of the large-sample distributions involved. These descriptions can be used to, for example, illustrate the “overoptimism” involved in applying the standard output from Cox regression analysis too simplistically, as when setting a 95% confidence interval for a parameter based on the AIC-selected model without taking into account the extra uncertainty associated with the model selection step. (See sec. 4 of HC 2003a for concrete illustrations of this.) The same phenomenon affects the significance levels of tests, where a tentative 5% test may actually have a much larger size.

So far we have not touched directly on application of the large-sample results to supplying “real” confidence intervals and “real” significance tests. This is not easy, in part because the required limit distributions of $\sqrt{n}(\hat{\mu} - \mu)$ are nonlinear mixtures of many normals. This is illustrated in Figure 1 for four estimation strategies for the case of the second focus estimand μ_2 in the Danish skin cancer survival study. Confidence intervals and tests with correct confidence levels and significance for large samples can be constructed based on assessment of these limit distributions in several ways (see, e.g., the “better confidence” recipe and the general discussion in HC 2003a, sec. 4).

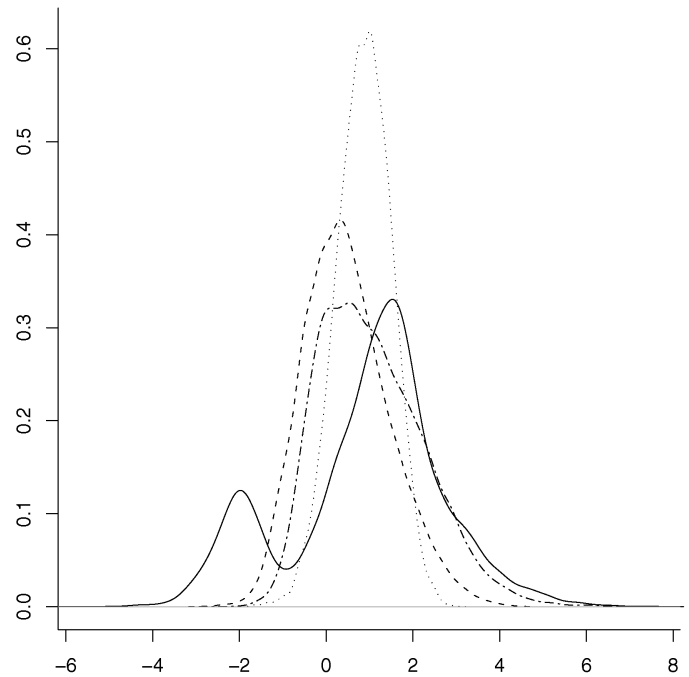


Figure 1. Results for the Second Focus Estimand μ_2 of Section 9.2, the Cumulative Hazard Rate Function for the Danish Skin Cancer Survival Data. Shown are densities for the limit distributions Λ for $\sqrt{n}(\hat{\mu}_2 - \mu_2)$ for four different estimators: the FIC method (—), the AIC method (---), the smoothed FIC method (-.-.-), and the smoothed AIC method (.....). The densities are computed as kernel estimates based on 10,000 simulations from the four appropriate versions of Λ as per Theorem 1 of Section 8, at the position $\hat{\eta} = \sqrt{n}\hat{\gamma}_{full}$ in the η parameter space.

APPENDIX: PROOFS

Proof of Lemma 1

We start from Taylor expansion

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} U_n(\hat{\beta}_S, \hat{\gamma}_S) \\ V_{n,S}(\hat{\beta}_S, \hat{\gamma}_S) \end{pmatrix} = \begin{pmatrix} U_n(\beta, 0) \\ V_{n,S}(\beta, 0) \end{pmatrix} + I_{n,S}(\beta, 0) \begin{pmatrix} \hat{\beta}_S - \beta \\ \hat{\gamma}_S \end{pmatrix},$$

which, with appropriate analysis of the error term involved, leads to

$$\begin{pmatrix} \sqrt{n}(\hat{\beta}_S - \beta) \\ \sqrt{n}\hat{\gamma}_S \end{pmatrix} = \{-I_{n,S}(\beta, 0)\}^{-1} \begin{pmatrix} \sqrt{n}U_n(\beta, 0) \\ \sqrt{n}V_{n,S}(\beta, 0) \end{pmatrix} + o_p(1).$$

Introduce the martingales

$$M_i(t) = N_i(t) - \int_0^t Y_i(u) \exp(\mathbf{x}_i^t \beta + \mathbf{z}_i^t \eta / \sqrt{n}) dH_0(u) \quad \text{for } i = 1, \dots, n.$$

These are orthogonal with variance functions $\langle M_i, M_i \rangle(t) = \int_0^t Y_i(u) \times \exp(\mathbf{x}_i^t \beta + \mathbf{z}_i^t \eta / \sqrt{n}) dH_0(u)$ (see, e.g., Andersen et al. 1993). We may then write

$$\begin{aligned} & \begin{pmatrix} \sqrt{n}U_n(\beta, 0) \\ \sqrt{n}V_{n,S}(\beta, 0) \end{pmatrix} \\ &= n^{-1/2} \sum_{i=1}^n \int_0^\tau \left\{ \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{i,S} \end{pmatrix} - E_{n,S}(u, \beta, 0) \right\} dM_i(u) \\ & \quad + n^{-1/2} \sum_{i=1}^n \int_0^\tau \left\{ \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{i,S} \end{pmatrix} - E_{n,S}(u, \beta, 0) \right\} \\ & \quad \times Y_i(u) \exp(\mathbf{x}_i^t \beta + \mathbf{z}_i^t \eta / \sqrt{n}) dH_0(u). \end{aligned}$$

The first term is an integral of a previsible function with respect to a martingale, and thus is itself a martingale evaluated at τ and with total

variance matrix

$$n^{-1} \sum_{i=1}^n \int_0^\tau \left\{ \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{i,S} \end{pmatrix} - \mathbf{E}_{n,S}(u, \boldsymbol{\beta}, \mathbf{0}) \right\} \left\{ \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{i,S} \end{pmatrix} - \mathbf{E}_{n,S}(u, \boldsymbol{\beta}, \mathbf{0}) \right\}^t \\ \times Y_i(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\eta} / \sqrt{n}) dH_0(u),$$

which is seen to converge in probability toward $\mathbf{J}_S$. After some algebra and analysis, the second term may be written as

$$n^{-1} \sum_{i=1}^n \int_0^\tau \left\{ \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{i,S} \end{pmatrix} - \mathbf{E}_{n,S}(u, \boldsymbol{\beta}, \mathbf{0}) \right\} \\ \times \mathbf{z}_i^t Y_i(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta}) dH_0(u) \boldsymbol{\eta} + O_p(n^{-1/2}),$$

which is seen to converge to the matrix with first p rows equal to $\mathbf{J}_{01} \boldsymbol{\eta}$ and the next $|S|$ rows equal to $\boldsymbol{\pi}_S \mathbf{J}_{11} \boldsymbol{\eta}$. This proves the lemma.

Proof of Lemma 2

With some Taylor approximation analysis, we find that $d\hat{H}_{0,S}(u)$ may be expressed as

$$\frac{\sum_{i=1}^n dN_i(u)}{nG_n^{(0)}(u, \hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S)} = \frac{\sum_{i=1}^n \{dM_i(u) + Y_i(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\eta} / \sqrt{n})\}}{nG_n^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})} \\ \times \left\{ 1 - \frac{\mathbf{G}_{n,S}^{(1)}(u, \boldsymbol{\beta}, \mathbf{0})^t}{G_n^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})} \begin{pmatrix} \hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta} \\ \hat{\boldsymbol{\gamma}}_S \end{pmatrix} \right\},$$

plus terms of order $O_p(n^{-1})$, which, after some further effort, becomes

$$\left[dH_0(u) + \frac{\sum_{i=1}^n dM_i(u)}{nG_n^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})} \right. \\ \left. + n^{-1/2} \frac{\sum_{i=1}^n Y_i(u) \exp(\mathbf{x}_i^t \boldsymbol{\beta}) \mathbf{z}_i^t dH_0(u)}{nG_n^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})} \boldsymbol{\eta} + o_p(n^{-1}) \right] \\ \times \left\{ 1 - \mathbf{E}_{n,S}(u, \boldsymbol{\beta}, \mathbf{0})^t \begin{pmatrix} \hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta} \\ \hat{\boldsymbol{\gamma}}_S \end{pmatrix} \right\}.$$

This leads to

$$n^{1/2} \{d\hat{H}_{0,S}(u) - dH_0(u)\} \\ = n^{-1/2} \frac{\sum_{i=1}^n dM_i(u)}{G_n^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})} - \mathbf{E}_n(u, \boldsymbol{\beta}, \mathbf{0})^t \left(\frac{\sqrt{n}(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta})}{\sqrt{n}\hat{\boldsymbol{\gamma}}_S} \right) dH_0(u) \\ + G_n^{(0)}(u, \boldsymbol{\beta}, \mathbf{0})^{-1} \mathbf{G}_{n,2}^{(1)}(u, \boldsymbol{\beta}, \mathbf{0})^t dH_0(u) \boldsymbol{\eta} + o_p(1) \\ \rightarrow_d dW(u) - \mathbf{e}(u, \boldsymbol{\beta}, \mathbf{0})^t \begin{pmatrix} \mathbf{B}_S \\ \mathbf{C}_S \end{pmatrix} dH_0(u) + \mathbf{e}_2(u, \boldsymbol{\beta}, \mathbf{0})^t \boldsymbol{\eta} dH_0(u),$$

which proves the lemma.

Proof of Lemma 3

It is helpful to translate $\mathbf{B}_S$ and $\mathbf{C}_S$ to other representations that better reveal the underlying bias and variance balance. Going back to the proof of Lemma 1, let us write

$$\begin{pmatrix} \sqrt{n} \mathbf{U}_n(\boldsymbol{\beta}, \mathbf{0}) \\ \sqrt{n} \mathbf{V}_n(\boldsymbol{\beta}, \mathbf{0}) \end{pmatrix} \rightarrow_d \begin{pmatrix} \mathbf{J}_{01} \boldsymbol{\eta} \\ \mathbf{J}_{11} \boldsymbol{\eta} \end{pmatrix} + \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix},$$

$$\text{where } \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim N_{p+q}(\mathbf{0}, \mathbf{J}_{\text{full}}).$$

Next, let $\mathbf{V}' = \mathbf{J}^{10} \mathbf{U} + \mathbf{J}^{11} \mathbf{V} = \mathbf{K}(\mathbf{V} - \mathbf{J}_{10} \mathbf{J}_{00}^{-1} \mathbf{U})$. We may show that $\mathbf{V}' \sim N_q(\mathbf{0}, \mathbf{K})$, independently of $\mathbf{U} \sim N_p(\mathbf{0}, \mathbf{J}_{00})$. Further algebra leads to

$$\mathbf{B}_S = (\mathbf{J}^{00,S} \mathbf{J}_{01} + \mathbf{J}^{01,S} \boldsymbol{\pi}_S \mathbf{J}_{11}) \boldsymbol{\eta} + \mathbf{J}^{00,S} \mathbf{U} + \mathbf{J}^{01,S} \mathbf{V}_S \\ = \mathbf{J}_{00}^{-1} \mathbf{J}_{01} (\mathbf{I} - \boldsymbol{\Omega}_S) \boldsymbol{\eta} + \mathbf{J}_{00}^{-1} \mathbf{U} - \mathbf{J}_{00}^{-1} \mathbf{J}_{01} \boldsymbol{\pi}_S^t \mathbf{K}_S \boldsymbol{\pi}_S (\mathbf{V} - \mathbf{J}_{10} \mathbf{J}_{00}^{-1} \mathbf{U}) \\ = \mathbf{J}_{00}^{-1} \mathbf{J}_{01} (\mathbf{I} - \boldsymbol{\Omega}_S) \boldsymbol{\eta} + \mathbf{J}_{00}^{-1} \mathbf{U} - \mathbf{J}_{00}^{-1} \mathbf{J}_{01} \boldsymbol{\Omega}_S \mathbf{V}',$$

whereas, similarly,

$$\mathbf{C}_S = (\mathbf{J}^{10,S} \mathbf{J}_{01} + \mathbf{J}^{11,S} \boldsymbol{\pi}_S \mathbf{J}_{11}) \boldsymbol{\eta} + \mathbf{J}^{10,S} \mathbf{U} + \mathbf{J}^{11,S} \mathbf{V}_S \\ = \mathbf{K}_S \boldsymbol{\pi}_S \mathbf{K}^{-1} (\boldsymbol{\eta} + \mathbf{V}').$$

We may also use this in conjunction with Lemma 2 to derive an alternative expression for $A_S(t)$ that better identifies the bias and variance parts.

We are now in a better position to work with $\hat{\mu}_S$. It is not difficult to derive representation (12) from Lemmas 1 and 2 through the delta method. Using the expressions found for $\mathbf{B}_S$ and $\mathbf{C}_S$, we may isolate the bias and the random parts of Λ_S . The random part may, after some algebra, be expressed as

$$\left(\frac{\partial \mu}{\partial H_0} \right) W(t) + \left\{ \frac{\partial \mu}{\partial \boldsymbol{\beta}} - \frac{\partial \mu}{\partial H_0} \mathbf{F}_0(t) \right\}^t \mathbf{J}_{00}^{-1} \mathbf{U} + (\kappa - \omega)^t \boldsymbol{\Omega}_S \mathbf{V}',$$

with variance as indicated earlier. Similarly, the nonrandom parts may be collected together and expressed as $\mathbf{b}_S^t \boldsymbol{\eta}$, where we indeed find that $\mathbf{b}_S = (\mathbf{I} - \boldsymbol{\Omega}_S^t)(\omega - \kappa)$.

Proof of Lemma 4

An essential ingredient in our proof is that the earlier process convergence result $A_{n,S}(t) = \sqrt{n}\{\hat{H}_{0,S}(t) - H_0(t)\} \rightarrow_d A_S(t)$ from Lemma 2, implies that

$$\Gamma_{n,S}(u) = \sqrt{n}\{\hat{H}_{0,S}^{-1}(u) - H_0^{-1}(u)\} \rightarrow_d \Gamma_S(u) = -\frac{A_S(H_0^{-1}(u))}{h_0(H_0^{-1}(u))}$$

as a process in u , in the Skorokhod topology over each compact interval. That this is true follows from general inversion results and techniques presented and discussed by Doss and Gill (1992) and Burr and Doss (1993).

We now work with

$$\hat{\xi}_S - \xi_{\text{true}} = \hat{H}_{0,S}^{-1}(f(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S, \boldsymbol{\gamma}_{0,S^c})) - H_0^{-1}(f(\boldsymbol{\beta}, \boldsymbol{\eta} / \sqrt{n})) \\ = \hat{H}_{0,S}^{-1}(f(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S, \boldsymbol{\gamma}_{0,S^c})) - H_0^{-1}(f(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S, \boldsymbol{\gamma}_{0,S^c})) \\ + H_0^{-1}(f(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S, \boldsymbol{\gamma}_{0,S^c})) - H_0^{-1}(f(\boldsymbol{\beta}, \mathbf{0})) \\ + H_0^{-1}(f(\boldsymbol{\beta}, \mathbf{0})) - H_0^{-1}(f(\boldsymbol{\beta}, \boldsymbol{\eta} / \sqrt{n})),$$

which decomposes our $\Lambda_{n,S}$ into three different sources of variation. With proper Taylor expansion arguments, we find that

$$\Lambda_{n,S} = \Gamma_{n,S}(f(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S)) + (H_0^{-1})'(f(\boldsymbol{\beta}, \mathbf{0}))\{f(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\gamma}}_S) - f(\boldsymbol{\beta}, \mathbf{0})\} \\ - (H_0^{-1})'(f(\boldsymbol{\beta}, \mathbf{0}))\{f(\boldsymbol{\beta}, \boldsymbol{\eta} / \sqrt{n}) - f(\boldsymbol{\beta}, \mathbf{0})\} + o_p(1),$$

which, by Lemmas 1 and 2, must have a limit distribution with representation

$$\Lambda_S = \Gamma_S(f(\boldsymbol{\beta}, \mathbf{0})) + (H_0^{-1})'(f(\boldsymbol{\beta}, \mathbf{0})) \left\{ \left(\frac{\partial f}{\partial \boldsymbol{\beta}} \right)^t \mathbf{B}_S + \left(\frac{\partial f}{\partial \boldsymbol{\gamma}_S} \right)^t \mathbf{C}_S \right\} \\ - (H_0^{-1})'(f(\boldsymbol{\beta}, \mathbf{0})) \left(\frac{\partial f}{\partial \boldsymbol{\gamma}} \right)^t \boldsymbol{\eta} \\ = h_0(\xi_0)^{-1} \\ \times \left[-A_S(\xi_0) + \left\{ \left(\frac{\partial f}{\partial \boldsymbol{\beta}} \right)^t \mathbf{B}_S + \left(\frac{\partial f}{\partial \boldsymbol{\gamma}_S} \right)^t \mathbf{C}_S \right\} - \left(\frac{\partial f}{\partial \boldsymbol{\gamma}} \right)^t \boldsymbol{\eta} \right].$$

In particular, the factor $h_0(H_0^{-1}(f(\boldsymbol{\beta}, \mathbf{0})))$ enters each of the contributions here and will be of no consequence when the task is to compare limiting risks for different subset models S .

The statement of Lemma 4 follows from these results, using representations for $\mathbf{B}_S$ and $\mathbf{C}_S$ arrived at in the course of proving Lemma 3.

Proof of Theorem 1

There is joint convergence in distribution of all $(\Lambda_{n,S}, \hat{\eta}, \mathbf{w}_n(S | \hat{\eta}))$ to that of $(\Lambda_S, \mathbf{D}, \mathbf{w}(S | \mathbf{D}))$, as can be seen from previous proofs. Also, from representation (13), $\Lambda_S = \Lambda_0 + (\omega - \kappa)^t(\eta - \Omega_S \mathbf{D})$. The statement of the lemma follows from this. (Also see the corresponding discussion in HC 2003a,b.)

[Received November 2004. Revised October 2005.]

REFERENCES

- Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, Heidelberg: Springer-Verlag.
- Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large-Sample Study," *The Annals of Statistics*, 10, 1100–1120.
- Breiman, L. (2001), "Statistical Modeling: The Two Cultures," *Statistical Science*, 16, 199–231.
- Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618.
- Bunea, F., and McKeague, I. W. (2005), "Covariate Selection for Semiparametric Hazard Function Regression Models," *Journal of Multivariate Analysis*, 92, 186–204.
- Burnham, K. P., and Anderson, D. R. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.), New York: Springer-Verlag.
- Burr, D., and Doss, H. (1993), "Confidence Bands for the Median Survival Time as a Function of the Covariates in the Cox Model," *Journal of the American Statistical Association*, 88, 1330–1340.
- Cai, J., Fan, J., Li, R., and Zhou, H. (2005), "Variable Selection for Multivariate Failure Time Data," *Biometrika*, 92, 303–316.
- Claeskens, G., and Hjort, N. L. (2003), "The Focused Information Criterion" (with discussion), *Journal of the American Statistical Association*, 98, 900–916.
- (2004), "Goodness of Fit via Nonparametric Likelihood Ratios," *Scandinavian Journal of Statistics*, 31, 487–513.
- Clyde, M. (1999), "Bayesian Model Averaging and Model Search Strategies" (with discussion), in *Bayesian Statistics 6*, eds. J. M. Bernardo, A. P. Dawid, J. O. Berger, and A. F. Smith, Oxford, U.K.: Oxford University Press, pp. 157–185.
- Dabrowska, D. M., and Doksum, K. (1987), "Estimates and Confidence Intervals for Median and Mean Life in the Proportional Hazards Model," *Biometrika*, 74, 799–807.
- Doss, H., and Gill, R. D. (1992), "An Elementary Approach to Weak Convergence for Quantile Processes, With Applications to Censored Survival Data," *Journal of the American Statistical Association*, 87, 869–877.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499.
- Fan, J., and Li, R. (2002), "Variable Selection for Cox's Proportional Hazards Model and Frailty Model," *The Annals of Statistics*, 30, 74–99.
- Foster, D. P., and George, E. I. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975.
- Gill, R. D. (1984), "Understanding Cox's Regression Model: A Martingale Approach," *Journal of the American Statistical Association*, 79, 441–447.
- Gould, S. J. (1995), "The Median Isn't the Message," in *Adam's Navel and Other Essays*, New York: Penguin Books, pp. 15–21.
- Hjort, N. L. (1992), "On Inference in Parametric Survival Data Models," *International Statistical Review*, 40, 355–387.
- Hjort, N. L., and Claeskens, G. (2003a), "Frequentist Model Average Estimators" (with discussion), *Journal of the American Statistical Association*, 98, 879–899.
- (2003b), "Rejoinder," *Journal of the American Statistical Association*, 98, 938–945.
- Hjort, N. L., and Pollard, D. B. (1996), "Asymptotics for Minimisers of Convex Processes," statistical research report, University of Oslo.
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial" (with discussion), *Statistical Science*, 19, 382–417; revised version available at <http://www.stat.washington.edu/raftery/>.
- Leeb, H., and Pötscher, B. M. (2004), "Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator," Technical Report 2004dec-1, Yale University, Dept. of Statistics.
- (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.
- Longford, N. T. (2005), "Model Selection and Efficiency: Is 'Which Model...?' the Right Question? [editorial]," *Journal of the Royal Statistical Society, Ser. A*, 168, 469–472.
- Tibshirani, R. (1997), "The Lasso Method for Variable Selection in the Cox Model," *Statistics in Medicine*, 16, 385–395.
- Volinsky, C., Madigan, D., Raftery, A. E., and Kronmal, R. A. (1997), "Bayesian Model Averaging in Proportional Hazard Models: Predicting the Risk of a Stroke," *Applied Statistics*, 46, 443–448.
- Volinsky, C., and Raftery, A. E. (2000), "Bayesian Information Criterion for Censored Survival Models," *Biometrics*, 56, 256–262.
- Yang, Y. (2005), "Can the Strengths of AIC and BIC Be Shared? *Biometrika*, 92, 937–950.