34

Focused Information Criteria for the Linear Hazard Regression Model

Nils Lid Hjort

Department of Mathematics, University of Oslo, Oslo, Norway

Abstract: The linear hazard regression model developed by Aalen is becoming an increasingly popular alternative to the Cox multiplicative hazard regression model. There are no methods in the literature for selecting among different candidate models of this nonparametric type, however. In the present chapter a focused information criterion is developed for this task. The criterion works for each specified covariate vector, by estimating the mean squared error for each candidate model's estimate of the associated cumulative hazard rate; the finally selected model is the one with lowest estimated mean squared error. Averaged versions of the criterion are also developed.

Keywords and Phrases: Aalen's linear model, covariate selection, focused information criterion, hazard regression, model selection

34.1 Introduction: Which Covariates to Include?

We consider survival regression data of the usual form (T_i, δ_i, x_i) for individuals $i = 1, \ldots, n$, where x_i is a vector of say r covariates, among which one wishes to select those of highest relevance. Also, $T_i = \min\{T_i^0, C_i\}$ is the possibly censored life-length and $\delta_i = I\{T_i^0 < C_i\}$ the associated noncensoring indicator, in terms of underlying life-length T_i^0 and censoring time C_i for individual i.

Our framework is that of the linear hazard regression model introduced by Aalen (1980); see; for example, the extensive discussion in Andersen *et al.* (1993, Ch. 8) and Martinussen and Scheike (2006, Ch. 5), where the hazard rate for individual i may be represented as

$$h_i(u) = x_i^{\mathrm{t}}\alpha(u) = \sum_{j=1}^{r} x_{i,j}\alpha_j(u) \quad \text{for } i = 1, \ldots, n,$$

in terms of regressor functions $\alpha_1(u), \ldots, \alpha_r(u)$. These need to satisfy the requirement that the linear combination $x^{\mathrm{t}}\alpha(u)$ stays nonnegative for all x supported by the distribution of covariate vectors. In other words, the associated cumulative hazard function

$$H(t \,|\, x) = \int_0^t x^{\mathrm{t}}\alpha(u)\,\mathrm{d}u = x^{\mathrm{t}}A(u) = \sum_{j=1}^r x_j A_j(t) \tag{34.1}$$

is nondecreasing in t, for all x in the relevant covariate space; here we write $A_j(t) = \int_0^t \alpha_j(u)\,\mathrm{d}u$ for $j = 1, \ldots, r$.

Among questions discussed in this chapter is when we might do better with only a subset of the x covariates than with keeping them all. We focus specifically on the problem of estimating $H(t \,|\, x)$ of (34.1) well, for a specified individual carrying his given covariate information x. The full-model estimator

$$\widehat{H}(t \,|\, x) = \widehat{H}_{\mathrm{full}}(t \,|\, x) = x^{\mathrm{t}}\widehat{A}(t) = \sum_{j=1}^r x_j \widehat{A}_j(t) \tag{34.2}$$

is one option, using the familiar Aalen estimators for $A_1, \ldots, A_r$ in the full model, keeping all covariates on board. Pushing some covariates out of the model leads to competing estimators of the type

$$\widetilde{H}_I(t \,|\, x) = \sum_{j \in I} x_j \widetilde{A}_{I,j}(t), \tag{34.3}$$

where the index set I is a subset of $\{1, \ldots, r\}$, representing those covariates that are kept in the model, and where the $\widetilde{A}_{I,j}(t)$s for $j \in I$ are the Aalen estimators in the linear hazard rate model associated with the I covariates. Using $\widetilde{H}_I(t \,|\, x)$ instead of $\widehat{H}(t \,|\, x)$ will typically correspond to smaller variances but to modelling bias. Slightly more generally, bigger index sets I imply more variance but less modelling bias, and vice versa. Thus the task of selecting suitable covariates amounts to a statistical balancing game between sampling variability and bias.

In Section 34.2 we fix the framework and give proper definitions of full-model and submodel estimators. These are also expressed in terms of counting processes and at-risk processes. Links with martingale theory make it possible in Section 34.3 to accurately assess the bias and variance properties associated with a given candidate model. This is followed up in Section 34.4 by explicit methods for estimating bias and variance from the data. The focused information criterion (FIC) introduced in Section 34.5 acts by estimating the risk associated with each candidate model's estimator of the cumulative hazard function; the model we suggest being used in the end is the one with the lowest estimated risk. Weighted versions are also put forward. In an extended version of the present work the use of the methods for real data and in some simulation

setups will be reported. This chapter ends with a list of concluding remarks in Section 34.7.

The brief introduction has so far taken model comparison as corresponding to accuracy of estimators of cumulative hazard rates $H(t\,|\,x)$. By a delta method argument this is also nearly equivalent to ranking models in terms of accuracy of estimates of survival probabilities $S(t\,|\,x) = \exp\{-H(t\,|\,x)\}$, where the estimates in question take the form

$$\widehat{S}_{\rm full}(t\,|\,x) = \prod_{[0,t]} \{1 - x^{\rm t}\,\widehat{A}(u)\} \quad \text{and} \quad \widetilde{S}_I(t\,|\,x) = \prod_{[0,t]} \Big\{1 - \sum_{j\in I} x_j\,{\rm d}\widetilde{A}_{I,j}(u)\Big\}.$$

[For details regarding notation for and properties of the product integral used on the right; see, for example, Andersen *et al.* (1993, Ch. II.6).] It is important to realise that a submodel I may work better than the full model, even if the submodel in question is not 'fully correct' as such; this is determined, among other aspects, by the sizes of the $\alpha_j(u)$ regressor functions that are left out of a model. This makes model selection different in spirit and operation than, for example, performing goodness-of-fit checks on all candidate models.

Aalen's linear hazard model is in many important respects different from Cox's proportional hazard model, also regarding the mathematical treatment of estimators and their properties; see Andersen *et al.* (1993, Ch. II.6). We note that focused information criteria and a general theory for model averaging estimators for the Cox model have been developed in Hjort and Claeskens (2006). Based on research in that and in the present chapter methods may be devised that can help select between 'the best Cox model' and 'the best Aalen model', in situations where that question is of relevance, but that theme is not pursued here.

34.2 Estimators in Submodels

This section properly defines the Aalen estimators $\widehat{A}$ and $\widetilde{A}_I$ involved in (34.2) and (34.3). It is convenient to define these in terms of the counting process and at-risk process

$$N_i(t) = I\{T_i \le t, \delta_i = 1\} \quad \text{and} \quad Y_i(u) = I\{T_i \ge u\}$$

for individuals $i = 1, \ldots, n$. We also need the martingales $M_i(t) = N_i(t) - \int_0^t Y_i(u)\,{\rm d}H_i(u)$, for which

$$ {\rm d}N_i(u) = Y_i(u)x_i^{\rm t}\,{\rm d}A(u) + {\rm d}M_i(u). \tag{34.4}$$

These are orthogonal and square integrable with variance processes

$$\langle M_i, M_i\rangle(t) = \int_0^t Y_i(u)h_i(u)\,{\rm d}u = \int_0^t Y_i(u)x_i^{\rm t}\,{\rm d}A(u). \tag{34.5}$$

In other words, $M_i(t)^2 - \langle M_i, M_i\rangle(t)$ is another zero-mean martingale, implying in particular that the mean of (34.5) is equal to the variance of $M_i(t)$.

Now introduce the $r \times r$-size matrix function

$$G_n(u) = n^{-1} \sum_{i=1}^{n} Y_i(u) x_i x_i^{\mathrm{t}}. \tag{34.6}$$

The Aalen estimator $\widehat{A} = (\widehat{A}_1, \ldots, \widehat{A}_r)^{\mathrm{t}}$ in the full model corresponds to

$$\mathrm{d}\widehat{A}(u) = G_n(u)^{-1} n^{-1} \sum_{i=1}^{n} x_i \, \mathrm{d}N_i(u),$$

with integrated version

$$\widehat{A}(t) = \int_0^t G_n(u)^{-1} n^{-1} \sum_{i=1}^{n} x_i \, \mathrm{d}N_i(u) \quad \text{for } t \geq 0. \tag{34.7}$$

This also defines $\widehat{H}_{\mathrm{full}}(t \,|\, x)$ of (34.2). It is assumed here that at least r linearly independent covariate vectors x_i remain in the risk set at time t, making the inverse of G_n well defined for all $u \leq t$; this event has probability growing exponentially quickly to 1 as sample size increases, under mild conditions.

To properly define the competitor $\widetilde{H}_I(t \,|\, x)$ of (34.3), we use the notation $x_I = \pi_I x$ for the vector of those x_j components for which $j \in I$, for each given subset I of $\{1, \ldots, r\}$. In other words, π_I is the projection matrix of size $|I| \times r$, with $|I|$ the number of covariates included in I. For the given I, we partition the G_n function into blocks,

$$G_n(u) = \begin{pmatrix} G_{n,00}(u), & G_{n,01}(u) \\ G_{n,10}(u), & G_{n,11}(u) \end{pmatrix},$$

where

$$G_{n,00}(u) = \pi_I G_n(u) \pi_I^{\mathrm{t}} = n^{-1} \sum_{i=1}^{n} Y_i(u) x_{i,I} x_{i,I}^{\mathrm{t}}$$

is of size $|I| \times |I|$, and $G_{n,11}(u)$ is of size $q \times q$ with $q = r - |I|$, and so on. The Aalen estimator for the vector of A_j functions where $j \in I$ is

$$\widetilde{A}_I(t) = \int_0^t G_{n,00}(u)^{-1} n^{-1} \sum_{i=1}^{n} x_{i,I} \, \mathrm{d}N_i(u).$$

These are those at work in (34.3).

Using (34.4) we may write

$$n^{-1} \sum_{i=1}^{n} x_{i,I} \, \mathrm{d}N_i(u) = n^{-1} \sum_{i=1}^{n} Y_i(u) x_{i,I} x_i^{\mathrm{t}} \, \mathrm{d}A(u) + n^{-1} \sum_{i=1}^{n} x_{i,I} \, \mathrm{d}M_i(u),$$

which further leads to

$$\begin{aligned} \mathrm{d}\widetilde{A}_I(u) \,=\, G_{n,00}(u)^{-1}\Big\{ & G_{n,00}(u)\,\mathrm{d}A_I(u) + G_{n,01}(u)\,\mathrm{d}A_{II}(u) \\ & + n^{-1}\sum_{i=1}^{n} x_{i,I}\,\mathrm{d}M_i(u)\Big\}, \end{aligned} \tag{34.8}$$

along with its integrated version. Here $II = I^c$ is the set of indexes not in I. This representation, basically in terms of a mean term plus martingale noise, is used in the next section to characterise means and variances of the (34.3) estimators. It again assumes that the G_n is invertible on $[0,t]$, an event having probability growing exponentially to 1 and therefore not disturbing the main analysis.

We remark that when the I model is used, then the Aalen estimator $\widetilde{A}_I(t)$ does not directly estimate A_I, but rather the function $A_I(t) + \int_0^t G_{00}^{-1} G_{01}\,\mathrm{d}A_{II}$.

34.3 Bias, Variance, and Mean Squared Error Calculations

In this section we develop useful approximations for the mean squared error of each of the (34.3) estimators $\widetilde{H}_I(t\,|\,x) = x_I^{\mathrm{t}}\widetilde{A}_I(t)$. We assume that the censoring variables $C_1,\ldots,C_n$ are i.i.d. with some survival distribution $C(u) = \Pr\{C_i \geq u\}$, and that they are independent of the lifetimes T_i^0; the case of no censoring corresponds to $C(u) = 1$ for all u. It is furthermore convenient to postulate that $x_1,\ldots,x_n$ stem from some distribution in the space of covariate vectors. These assumptions imply, for example, that the G_n function of (34.6) converges with increasing sample size, say

$$G_n(u) \to G(u) = \mathrm{E}_* Y(u) xx^{\mathrm{t}} = \mathrm{E}_* \exp\{-x^{\mathrm{t}}A(u)\}xx^{\mathrm{t}}\,C(u), \tag{34.9}$$

where E_* refers to expectation under the postulated covariate distribution. Also the mean function

$$\bar{G}_n(u) = \mathrm{E}\,G_n(u) = n^{-1}\sum_{i=1}^{n} p_i(u) x_i x_i^{\mathrm{t}}$$

converges to the same limit $G(u)$; here $p_i(u) = \mathrm{E}Y_i(u) = \exp\{-x_i^{\mathrm{t}}A(u)\}\,C(u)$. We finally assume that the $r \times r$-function $G(u)$ is invertible over the time observation window $u \in [0,\tau]$ of interest; this corresponds to $C(\tau)$ positive and to a nondegenerate covariate distribution. As in Section 34.2 there is a need to partition the $G(u)$ function into blocks $G_{00}(u), G_{01}(u)$, and so on; $G_{00}(u)$ has, for example, size $|I| \times |I|$. A similar remark applies to $\bar{G}_n(u)$.

Consider as in Section 34.1 a given individual with covariate information x. From representation (34.8),

$$\begin{aligned} x_I^{\mathrm{t}}\,\mathrm{d}\widetilde{A}_I(u) &= x_I^{\mathrm{t}}\,\mathrm{d}A_I(u) + x_I^{\mathrm{t}}G_{n,00}(u)^{-1}G_{n,01}(u)\,\mathrm{d}A_{II}(u) \\ &\quad + n^{-1/2}x_I^{\mathrm{t}}G_{n,00}(u)^{-1}\,\mathrm{d}V_{n,I}(u) \\ &= x^{\mathrm{t}}\,\mathrm{d}A(u) + b_{I,n}(u)^{\mathrm{t}}\,\mathrm{d}A_{II}(u) + n^{-1/2}x_I^{\mathrm{t}}G_{n,00}(u)^{-1}\,\mathrm{d}V_{n,I}(u), \end{aligned}$$

in which V_n is the r-dimensional martingale process with increments

$$\mathrm{d}V_n(u) = n^{-1/2}\sum_{i=1}^n x_i\,\mathrm{d}M_i(u), \tag{34.10}$$

whereas $b_{I,n}$, defined by

$$b_{I,n}(u) = G_{n,10}(u)G_{n,00}(u)^{-1}x_I - x_{II}, \tag{34.11}$$

can be seen as a bias function (omitting at the moment x in the notation for this function). Its dimension is $q = r - |I|$. This leads to the representation

$$\sqrt{n}\{x_I^{\mathrm{t}}\widetilde{A}_I(t) - x^{\mathrm{t}}A(t)\} = \sqrt{n}\int_0^t b_{I,n}^{\mathrm{t}}\,\mathrm{d}A_{II} + x_I^{\mathrm{t}}\int_0^t G_{n,00}^{-1}\,\mathrm{d}V_{n,I}. \tag{34.12}$$

The second term is a zero-mean martingale whereas the first term is a bias term, stemming from using model I that does not include all the components. We use (34.12) to develop good approximations to

$$\mathrm{mse}_n(I) = \mathrm{mse}_n(I,t) = n\,\mathrm{E}\{\widetilde{H}_I(t\,|\,x) - H(t\,|\,x)\}^2,$$

the normalised mean squared error of the (34.3) estimator. We treat the covariate vectors $x_1,\ldots,x_n$ as given; that is, our approximations are expressed directly in terms of these.

In view of the assumptions made in the beginning of this section, a first-order approximation to the mean of (34.12) is $\sqrt{n}\int_0^t \bar{b}_{I,n}^{\mathrm{t}}\,\mathrm{d}A_{II}$, because the second term has zero mean; here $\bar{b}_{I,n}(u) = \bar{G}_{n,10}(u)\bar{G}_{n,00}(u)^{-1}$. Also, $\int_0^t b_{I,n}^{\mathrm{t}}\,\mathrm{d}A_{II}$ and $\int_0^t \bar{b}_{I,n}^{\mathrm{t}}\,\mathrm{d}A_{II}$ are both close to the limit $\int_0^t b_I^{\mathrm{t}}\,\mathrm{d}A_{II}$, with high probability for large n, where $b_I(u) = G_{10}(u)G_{00}^{-1}x_I - x_{II}$.

To study the second term of (34.12), note that V_n of (34.10) is a zero-mean martingale with variance process $\langle V_n, V_n\rangle(t) = J_n(t)$, with $r\times r$-matrix increments

$$\mathrm{d}J_n(u) = n^{-1}\sum_{i=1}^n Y_i(u)x_i x_i^{\mathrm{t}}\,x_i^{\mathrm{t}}\,\mathrm{d}A(u).$$

There is a well-defined limit function $J(u)$ with increments

$$\mathrm{d}J(u) = \mathrm{E}_* Y(u)xx^{\mathrm{t}}\,x^{\mathrm{t}}\,\mathrm{d}A(u) = \mathrm{E}_*\exp\{-x^{\mathrm{t}}A(u)\}xx^{\mathrm{t}}\,x^{\mathrm{t}}\,\mathrm{d}A(u)\,C(u)$$

under the conditions stated above. Thus V_n converges in distribution to a Gaussian martingale V with increments $\mathrm{d}V(u)$ having zero mean and variance matrix $\mathrm{d}J(u)$. It also follows that the second term of (34.12) converges in distribution,

$$x_I^{\mathrm{t}} \int_0^t G_{n,00}^{-1}\,\mathrm{d}V_{n,I} \to_d x_I^{\mathrm{t}} \int_0^t G_{00}^{-1}\,\mathrm{d}V_I,$$

which is normal with variance

$$\mathrm{var}(I,t) = x_I^{\mathrm{t}} \int_0^t G_{00}^{-1}\,\mathrm{d}J_{00}\,G_{00}^{-1} x_I.$$

The integral here is defined in the appropriate and natural Riemannian sense, and is also equivalent to a finite sum of ordinary integrals, found by writing out the quadratic form.

The first term of (34.12) is essentially nonrandom when compared with the second term. A more formal statement can be put forward in a framework of local asymptotic neighbourhoods, where $\mathrm{d}A_{II}(u) = \mathrm{d}D(u)/\sqrt{n}$, say; in this case,

$$\sqrt{n}\{\widetilde{H}_I(t\,|\,x) - H(t\,|\,x)\} \to_d \int_0^t b(u)^{\mathrm{t}}\,\mathrm{d}D(u) + \mathrm{N}(0, \mathrm{var}(I,t)).$$

Our main use of these considerations is the approximation to the normalised mean squared error;

$$\mathrm{mse}_n(I,t) \doteq \mathrm{sqb}(I,t) + \mathrm{var}(I,t), \tag{34.13}$$

where $\mathrm{var}(I,t)$ is defined above and

$$\mathrm{sqb}(I,t) = n\Big(\int_0^t \bar{b}_{I,n}^{\mathrm{t}}\,\mathrm{d}A_{II}\Big)^2.$$

Remark There are often situations where it pays to exclude some covariates, even though their associated $\alpha_j(u)$ functions are nonzero. This is a consequence of the squared bias versus variance balancing game. For example, a submodel I is better than the full set, for the given covariate x, if $\mathrm{sqb}(I,t) + \mathrm{var}(I,t) \le 0 + \mathrm{var}(\mathrm{full},t)$, which translates to

$$n\Big\{\int_0^t (G_{10}G_{00}^{-1}x_I - x_{II})^{\mathrm{t}}\,\mathrm{d}A_{II}\Big\}^2 \le x^{\mathrm{t}} \int_0^t G^{-1}\,\mathrm{d}J\,G^{-1}\,x - x_I^{\mathrm{t}} \int_0^t G_{00}^{-1}\,\mathrm{d}J_{00}\,G_{00}^{-1}\,x_I.$$

This effectively describes a 'tolerance radius' around a given model, inside which the model is preferable to the full model, even when not perfectly valid. The inequality says that a certain linear combination of the $\alpha_j(u)$ functions for $j \notin I$ should not be too big, compared also to the sample size; for large n even small biases are costly, and the full model becomes preferable.

34.4 Estimating the Risks

We have seen that each candidate model I has an associated risk $\text{mse}_n(I,t)$ of (34.13) when estimating the cumulative hazard function using $\widetilde{H}_I(t\,|\,x)$. Here we deal with the consequent task of estimating these risk quantities from data.

For the variance part we use

$$\widehat{\text{var}}(I,t) = x_I^{\text{t}} \int_0^t G_{n,00}^{-1}(u)\,\text{d}\widehat{J}_{n,00}(u)\,G_{n,00}(u)^{-1}\,x_I,$$

wherein

$$\text{d}\widehat{J}_n(u) = n^{-1}\sum_{i=1}^n Y_i(u)x_i x_i^{\text{t}}\,x_i^{\text{t}}\text{d}\widehat{A}(u),$$

engaging the full-model Aalen estimator. The $|I|\times|I|$ block used for the variance estimation is $\pi_I\,\text{d}\widehat{J}_n(u)\pi_I^{\text{t}}$.

For the squared bias part, consider in general terms the quantity β^2, where $\beta = \int_0^t g^{\text{t}}\,\text{d}A_{II}$, for a specified q-dimensional function g; again, $q = r - |I|$. Considering $\widehat{\beta} = \int_0^t g^{\text{t}}\,\text{d}\widehat{A}_{II}$, employing the II part of the full-model Aalen estimator, we have

$$\text{E}\,\widehat{\beta} \doteq \beta \quad\text{and}\quad \text{Var}\,\widehat{\beta} \doteq n^{-1}\int_0^t g(u)^{\text{t}}\,\text{d}Q(u)\,g(u),$$

from results above, where we write

$$\text{d}Q(u) = \{G(u)^{-1}\,\text{d}J(u)\,G(u)^{-1}\}_{11}$$

for the lower right-hand $q\times q$ block of the matrix within brackets, the block associated with subset $II = I^c$. Thus $\text{E}\,\widehat{\beta}^2 \doteq \beta^2 + n^{-1}\int_0^t g^{\text{t}}\,\text{d}Q\,g$, in its turn leading to the natural and nearly unbiased estimator

$$\Big(\int_0^t g^{\text{t}}\,\text{d}\widehat{A}_{II}\Big)^2 - n^{-1}\int_0^t g(u)^{\text{t}}\,\text{d}\widehat{Q}_n(u)\,g(u)$$

for β^2, where

$$\text{d}\widehat{Q}_n(u) = \pi_{II}\{G_n(u)^{-1}\,\text{d}\widehat{J}_n(u)\,G_n(u)^{-1}\}\pi_{II}^{\text{t}}$$

is the empirical counterpart to $\text{d}Q(u)$.

These considerations lead to the risk estimator

$$\widehat{R}(I,t) = \widehat{\text{mse}}_n(I,t) = \max\{\widehat{\text{sqb}}(I,t),0\} + x_I^{\text{t}}\int_0^t G_{n,00}^{-1}\,\text{d}\widehat{J}_{n,00}\,G_{n,00}^{-1}\,x_I,$$

where

$$\widehat{\text{sqb}}(I,t) = n\Big(\int_0^t b_{I,n}^{\text{t}}\,\text{d}\widehat{A}_{II}\Big)^2 - \int_0^t b_{I,n}^{\text{t}}\,\text{d}\widehat{Q}_n\,b_{I,n}.$$

34.5 The FIC and the Weighted FIC

Here we show how risk estimation methods developed above lead to natural information criteria for model selection.

The first such is a *focused information criterion* that works for a given individual and a given time point at which we wish optimal precision for her survival probability estimate. For the given covariate x and time point t we calculate

$$\mathrm{FIC} = \mathrm{FIC}(I, x, t) = \max\{\widehat{\mathrm{sqb}}(I, x, t), 0\} + \widehat{\mathrm{var}}(I, x, t) \tag{34.14}$$

for each candidate model I, where

$$\begin{aligned}
\widehat{\mathrm{sqb}}(I, x, t) &= n\Big(\int_0^t b_{I,n}^{\mathrm{t}}\, \mathrm{d}\widehat{A}_{II}\Big)^2 - \int_0^t b_{I,n}^{\mathrm{t}}\, \mathrm{d}\widehat{Q}_n\, b_{I,n}, \\
\widehat{\mathrm{var}}(I, x, t) &= x_I^{\mathrm{t}} \int_0^t G_{n,00}^{-1}\, \mathrm{d}\widehat{J}_{n,00}\, G_{n,00}^{-1}\, x_I.
\end{aligned}$$

We note that $b_{I,n}(u)$ of (34.11) depends on x and that the submatrices $G_{n,00}$ and so on of (34.9) depend on I. In the end one selects the model with smallest value of the FIC score number.

Note that FIC is sample-size dependent. In a situation with a given amount of nonzero bias $\int_0^t \bar{b}_I^{\mathrm{t}}\, \mathrm{d}A_{II}$, the $\widehat{\mathrm{sqb}}$ component of FIC will essentially increase with n, whereas the variance component remains essentially constant. This goes to show that the best models will tolerate less and less bias as n increases, and for sufficiently large n only the full model (which has zero modelling bias) will survive FIC scrutiny.

There are various variations on the FIC above. For a given individual who has survived up to time t_1 it is the conditional survival probabilities

$$\Pr\{T^0 \geq t_2 \,|\, T^0 \geq t_1, x\} = \exp[-\{H(t_2 \,|\, x) - H(t_1 \,|\, x)\}]$$

that are of interest. The development and formulae above can be repeated mutatis mutandis with a given interval $[t_1, t_2]$ replacing $[0, t]$. This gives a machinery for selecting models that yield optimal estimation precision for conditional survival probabilities. It will also be useful in many applications to monitor FIC scores for important candidate models in terms of a 'gliding time window', say $[t - \delta, t + \delta]$; successful models should then have good FIC scores across time. We stress that it is not a paradox that one model might be particularly good at explaining the survival mechanisms involved for short life-lengths, whereas another model might be much better for understanding the survival of the longer life-lengths. Our FIC takes this on board, and makes an explicit model recommendation for each given time interval of interest.

Suppose now that a model is called for that works well in an average sense across a given set of (x,t) values, as opposed to a given (x,t). Consider in general terms

$$\mathcal{E}_n(I) = n\int \{\widetilde{H}_I(t\,|\,x) - H(t\,|\,x)\}^2\,\mathrm{d}w(t,x),$$

where $w(t,x)$ is a weight measure in the (x,t) space. This could, for example, take the form

$$\mathcal{E}_n(I) = (1/K)\sum_{j=1}^{K} n\{\widetilde{H}_I(t\,|\,x_j) - H(t\,|\,x_j)\}^2, \tag{34.15}$$

averaging across given covariate vectors $x_1,\ldots,x_K$. From (34.12), the random loss incurred using I is

$$\mathcal{E}_n(I) = \int \Big\{\sqrt{n}\int_0^t b_{I,n}(u,x)^{\mathrm{t}}\,\mathrm{d}A_{II}(u) + x_I^{\mathrm{t}}\int_0^t G_{n,00}(u)^{-1}\,\mathrm{d}V_{n,I}(u)\Big\}^2\,\mathrm{d}w(t,x),$$

writing now

$$b_{I,n}(u,x) = G_{n,10}(u)G_{n,00}(u)^{-1}x_I - x_{II}$$

with explicit mention of x in the notation.

Its associated risk, the expected loss, is by previous efforts closely approximated by the w-weighted risk

$$R_n(I) = \mathrm{E}\int \Big[n\Big\{\int_0^t b_{I,n}(u,x)^{\mathrm{t}}\,\mathrm{d}A_{II}(u)\Big\}^2 + x_I^{\mathrm{t}}\int_0^t G_{n,00}^{-1}\,\mathrm{d}J_{n,00}\,G_{n,00}^{-1}\,x_I\Big]\,\mathrm{d}w(t,x).$$

We estimate the w-weighted squared bias and w-weighted variance contributions in turn. Define

$$\begin{aligned}\widehat{\text{w-sqb}}(I) \;=\; & n\int\Big\{\int_0^t b_{I,n}(u,x)^{\mathrm{t}}\,\mathrm{d}\widehat{A}_{II}(u)\Big\}^2\,\mathrm{d}w(t,x)\\ & - \int\int_0^t b_{I,n}(u,x)^{\mathrm{t}}\,\mathrm{d}\widehat{Q}_n(u)\,b_{I,n}(u,x)\,\mathrm{d}w(t,x),\end{aligned}$$

which is an approximately unbiased estimator of the w-weighted squared bias term; and

$$\text{w-}\widehat{\text{var}}(I) = \int \widehat{\text{var}}(I,x,t)\,\mathrm{d}w(t,x).$$

Our wFIC score, to be computed for each candidate model, is

$$\text{wFIC}(I) = \max\{\widehat{\text{w-sqb}}(I),0\} + \text{w-}\widehat{\text{var}}(I). \tag{34.16}$$

Again, in the end the model achieving the lowest wFIC score is selected. This scheme in particular gives rise to an algorithm associated with the (34.15) loss, weighting evenly across a finite set of covariate vectors.

A special case worth recording is when t is fixed and w describes the covariate distribution. It is unknown, but may be approximated with the empirical distribution of covariates $x_1, \ldots, x_n$. This leads to wFIC(I) as in (34.16) with

$$\begin{aligned} \text{w-}\widehat{\text{var}}(I) &= n^{-1} \sum_{i=1}^{n} \widehat{\text{var}}(I, x_i, t) \\ &= \text{Tr}\Big\{ \Big(\int_0^t G_{n,00}^{-1} \, \mathrm{d}\widehat{J}_{n,00} \, G_{n,00}^{-1} \Big) \Big(n^{-1} \sum_{i=1}^{n} x_{i,I} x_{i,I}^{\mathrm{t}} \Big) \Big\}, \end{aligned}$$

whereas $\text{w-}\widehat{\text{sqb}}(I)$ may be written

$$\sum_{i=1}^{n} \{ x_{i,I}^{\mathrm{t}} \widehat{B}_I(t) - x_{i,II}^{\mathrm{t}} \widehat{A}_{II}(t) \}^2 - n^{-1} \sum_{i=1}^{n} \int_0^t b_{I,n}(u, x_i)^{\mathrm{t}} \, \mathrm{d}\widehat{Q}_n(u) \, b_{I,n}(u, x_i),$$

where

$$\widehat{B}_I(t) = \int_0^t G_{n,00}(u)^{-1} G_{n,01}(u) \, \mathrm{d}\widehat{A}_{II}(u).$$

Remark Note that the wFIC method as defined here is subtly but crucially different from simply w-weighting of the individual pointwise FIC scores, regarding how the truncation of the squared bias estimate is carried out. In (34.16), the truncation to achieve nonnegativity of the estimate takes place after the w-weighting, making it different from w-weighting the collection of truncated sqb(I, x, t) terms. See in this connection also Claeskens and Hjort (2007).

34.6 Exact Risk Calculations

In the previous sections we were able to (i) develop formulae for risk functions and (ii) construct estimators for these. This led to model selection methods that may be used in any given application. The present section has a different aim, namely that of providing classes of case studies where the risk function formulae can be computed explicitly, thereby establishing a fair testing ground for model selection and model averaging methods. For reasons of space we are content to derive certain formulae under certain conditions, for biases and variances; these may then be used to form concrete illustrations and test cases that for reasons of space cannot be reported on in the present chapter.

Assume that the components $x_1, \ldots, x_r$ of the covariate vector x are distributed independently of each other, with Laplace transforms $\text{E}_* \exp(-\theta_j x_j) = \exp\{-M_j(\theta_j)\}$, say. Then

$$\text{E}_* \exp(-\theta^{\mathrm{t}} x) = \exp\{-M_1(\theta_1) - \cdots - M_r(\theta_r)\},$$

from which follows, taking second-order derivatives with respect to the θ components, that

$$\mathrm{E}_* \exp(-\theta^{\mathrm{t}}x)x_jx_k = \exp\Big\{-\sum_{l=1}^r M_l(\theta_l)\Big\}\{-M_j''(\theta_j)\delta_{j,k} + M_j'(\theta_j)M_k'(\theta_k)\},$$

in terms of first- and second-order derivatives of the M_j functions. This implies that the $r \times r$ limit function G of (34.9) may be expressed as

$$G(u) = f(u)\{D(u) + z(u)z(u)^{\mathrm{t}}\}C(u).$$

Here $f(u) = \exp\{-\sum_{l=1}^r M_l(A_l(u))\}$; $D(u)$ is the diagonal matrix with elements $D_j(u) = -M_j''(A_j(u))$; and $z(u)$ is the vector with elements $z_j(u) = M_j'(A_j(u))$. For a candidate set I of covariates to include, the blocks of $G(u)$ can be read off from

$$G(u) = f(u)C(u)\Big\{\begin{pmatrix} D_0 & 0 \\ 0 & D_1 \end{pmatrix} + \begin{pmatrix} z_0 \\ z_1 \end{pmatrix}\begin{pmatrix} z_0 \\ z_1 \end{pmatrix}^{\mathrm{t}}\Big\},$$

where D_0 and D_1 have components $D_j(u)$ where, respectively, $j \in I$ and $j \notin I$, and similarly z_0 and z_1 have components $z_j(u)$ where $j \in I$ and $j \notin I$. In particular,

$$G_{00}(u) = f(u)C(u)(D_0 + z_0z_0^{\mathrm{t}}) \quad \text{and} \quad G_{01}(u) = f(u)C(u)z_0z_1^{\mathrm{t}},$$

leading in turn, via the matrix inversion formula

$$(D_0 + z_0z_0^{\mathrm{t}})^{-1} = D_0^{-1} - \frac{1}{1 + z_0^{\mathrm{t}}D_0^{-1}z_0}D_0^{-1}z_0z_0^{\mathrm{t}}D_0^{-1},$$

to a formula for $G_{00}(u)^{-1}G_{01}(u)$ and then to

$$\begin{aligned} b_I(u) &= G_{10}(u)G_{00}(u)^{-1}x_I - x_{II} \\ &= z_1z_0^{\mathrm{t}}\Big(D_0^{-1} - \frac{1}{1 + z_0^{\mathrm{t}}D_0^{-1}z_0}D_0^{-1}z_0z_0^{\mathrm{t}}D_0^{-1}\Big)x_I - x_{II} \\ &= z_1\frac{z_0^{\mathrm{t}}D_0^{-1}x_I}{1 + z_0^{\mathrm{t}}D_0^{-1}z_0} - x_{II}. \end{aligned}$$

Assume for a concrete example that $x_j \sim \text{gamma}(a_j, b_j)$ for $j = 1, \ldots, r$, for which the Laplace transforms are $\{b_j/(b_j+\theta_j)\}^{a_j}$ with $M_j(\theta_j) = a_j\log(1+\theta_j/b_j)$. Then

$$M_j'(\theta_j) = \frac{\xi_j}{1 + \theta_j/b_j} \quad \text{and} \quad M_j''(\theta_j) = -\frac{\xi_j/b_j}{(1 + \theta_j/b_j)^2},$$

with $\xi_j = \mathrm{E}_*x_j = a_j/b_j$. This yields a bias function $b_I(u)$ with components

$$b_{I,j}(u) = \frac{g_I(u)}{1 + \sum_{j\in I} b_j\xi_j}\frac{\xi_j}{1 + A_j(u)/b_j} - x_j \quad \text{for } j \in II = I^c,$$

where $g_I(u) = \sum_{j\in I}\{b_j + A_j(u)\}x_j$. It follows that the important bias component of (34.12) may be written

$$\sqrt{n}\int_0^t b_I^{\mathrm{t}}\,\mathrm{d}A_{II} = \sqrt{n}\Big\{\int_0^t \frac{g_I(u)}{1+\sum_{j\in I} b_j\xi_j}\sum_{j\in II}\frac{\xi_j\alpha_j(u)}{1+A_j(u)/b_j}\,\mathrm{d}u - x_{II}^{\mathrm{t}}A_{II}(t)\Big\}.$$

These bias functions are easily computed and displayed, for given covariate distributions and given hazard regression functions.

To handle the variance part of (34.13) we need an explicit formula for $\mathrm{d}J(u)$ and then for

$$G_{00}^{-1}\,\mathrm{d}J_{00}\,G_{00}^{-1} \quad\text{and}\quad G(u)^{-1}\,\mathrm{d}J(u)\,G(u)^{-1}.$$

We start with

$$\mathrm{E}_*\exp(-s\theta^{\mathrm{t}}x)x_jx_k = \exp\Big\{-\sum_{l=1}^r M_l(s\theta_l)\Big\}\{-M_j''(s\theta_j)\delta_{j,k} + M_j'(s\theta_j)M_k'(s\theta_k)\},$$

and then take the derivative w.r.t. s, and set $s=1$ in the resulting equations. This yields

$$\begin{aligned}\mathrm{E}_*\exp\{-\theta^{\mathrm{t}}x)x_jx_k\,\theta^{\mathrm{t}}x = f^*(\theta)[&\{M_j'''(\theta_j)\theta_j - g^*(\theta)M_j''(\theta_j)\}\delta_{j,k}\\ &- M_j'(\theta_j)M_k''(\theta_k)\theta_k - M_j''(\theta_j)M_k'(\theta_k)\theta_j\\ &+ g^*(\theta)M_j'(\theta_j)M_k'(\theta_k)],\end{aligned}$$

where

$$f^*(\theta) = \exp\Big\{-\sum_{l=1}^r M_l(\theta_l)\Big\} \quad\text{and}\quad g^*(\theta) = \sum_{l=1}^r M_l'(\theta_l)\theta_l.$$

Let now $A_j(t) = \alpha_j t$ for $j = 1,\ldots,r$; that is, the α_j regressor functions are taken constant. The above leads with some further work to a formula for

$$\mathrm{E}_*\exp\{-x^{\mathrm{t}}A(u)\}xx^{\mathrm{t}}\,x^{\mathrm{t}}\,\mathrm{d}A(u) = f(u)\{E(u)+F(u)\}\,\mathrm{d}u,$$

where the $E(u)$ and $F(u)$ matrix functions are described below; also, $f(u) = \exp\{-\sum_{l=1}^r M_l(A_l(u))\}$ is as for the bias calculations above. The $E(u)$ is diagonal with elements

$$E_j(u) = M_j'''(A_j(u))\alpha_j - g(u)M_j''(A_j(u)),$$

where $g(u) = \sum_{l=1}^r M_l'(A_l(u))\alpha_l$. Next, $F(u)$ has (j,k) element

$$\begin{aligned}-M_j'(A_j(u))M_k''(A_k(u))\alpha_k \;-\; &M_j''(A_j(u))M_k'(A_k(u))\alpha_j\\ &+ g(u)M_j'(A_j(u))M_k'(A_k(u)).\end{aligned}$$

These results may be used to compute the variance terms

$$x_I^{\mathrm{t}} \int_0^t G_{00}^{-1}\,\mathrm{d}J_{00}\,G_{00}^{-1}\,x_I$$

and thereby the mean squared errors for different candidate models. These formulae may in particular be used for the case mentioned earlier, with independent gamma(a_j, b_j) distribution for the x_j components, and for which $M_j'''(\theta_j) = 2(\xi_j/b_j^2)/(1+\theta_j/b_j)^3$.

Various concrete illustrations may now be given, for the specific case of independent gamma distributed covariates, to exhibit and examine various aspects and issues involved in model selection and model averaging. These relate in various ways to modelling bias versus estimation variance. We may, for example, show that when $\alpha_j(u)$s are small in size, then it may be best not to include these in the selected model, depending also on the sizes of x_I and x_{II}. We would also be able to illustrate how the complexity of the best model increases with higher sample size, and how the qualitative results depend on the relative spread of the distributions of covariates.

34.7 Concluding Remarks

Here we offer some concluding comments, some pointing to natural extensions of the material and methods we have presented above.

1. In a planned extended version of this chapter space will be given to analysis of a real dataset and to instructive simulation setups.

2. We have throughout used 'vanilla weights' for the Aalen estimators $\widehat{A}$ of (34.7). With more sophisticated weighting the estimator

$$\widehat{A}(t,k) = \int_0^t \Big\{ n^{-1} \sum_{i=1}^n Y_i(u) k_i(u) x_i x_i^{\mathrm{t}} \Big\}^{-1} n^{-1} \sum_{i=1}^n x_i k_i(u)\,\mathrm{d}N_i(u)$$

   may perform slightly better; see Huffer and McKeague (1991). Also for such schemes a FIC and wFIC methodology may be developed, generalising methods given in the present chapter.

3. A local asymptotic framework may be put up for the Aalen model, similar in spirit to that employed in Hjort and Claeskens (2003) and Claeskens and Hjort (2003) for purely parametric models. Here one would use hazard rates

$$h_i(u) = \sum_{j=1}^p x_{i,j}\alpha_j(u) + \sum_{j=1}^q z_{i,j}\delta_j(u)/\sqrt{n},$$

with $x_{i,j}$s protected covariates considered important to include in all candidate models, and $z_{i,j}$s the potentially discardable ones. A precise asymptotic description may now be given of all limiting risk functions, in terms of the $\delta_1, \ldots, \delta_q$ functions.

4. A fair question to ask is the behaviour of the final estimator, say

$$H^*(t \,|\, x) = \widetilde{H}_{\widehat{I}}(t \,|\, x),$$

where $\widehat{I}$ is the data-dependent set of finally included covariates. This is a complicated question without any easy answer. Inside the local asymptotic framework of (3), methods of Hjort and Claeskens (2003) may be used to describe the limit distribution of $\sqrt{n}\{H^*(t \,|\, x) - H(t \,|\, x)\}$, in terms of a nonlinear mixture of biased normals. This also opens the door to general model average strategies, as opposed to limiting inference methods to those that rely on deciding on only one model.

5. We have developed machinery for answering the question, "Should covariate j be included in the nonparametric Aalen model, or not?". More ambitiously and more laboriously, one can give not only two but three potential outcomes for each covariate: it might be excluded; it might be included nonparametrically; or it might be included parametrically. The latter possibility refers for example, to the model where $\alpha_j(u)$ is constant; see McKeague and Sasieni (1994) for treatment of such models. Again a FIC and a wFIC apparatus may be developed, requiring, however, more mathematical vigour.

Acknowledgements

This chapter reports on work carried out in the stimulating environment of the Centre of Advanced Study at the Norwegian Academy of Science and Letters, where Ørnulf Borgan and Odd Aalen organised a research group on Statistical Analysis of Complex Event History Data. Conversations with Axel Gandy were particularly fruitful.

References

1. Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes, In *Mathematical Statistics and Probability Theory* (Eds., W. Klonecki, A. Kozek, and J. Rosinski). *Proceedings of the 6th International Conference*, Wisla, Poland, pp. 1–25.

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

3. Claeskens, G. and Hjort, N. L. (2003). The focused information criterion [with discussion], *Journal of the American Statistical Association*, **98**, 900–916 and 938–945.

4. Claeskens, G. and Hjort, N. L. (2008). Minimising average risk in regression models, *Econometric Theory*, to appear.

5. Hjort, N. L. and Claeskens, G. (2003). Frequentist average estimators [with discussion], *Journal of the American Statistical Association*, **98**, 879–899 and 938–945.

6. Hjort, N. L. and Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model, *Journal of the American Statistical Association*, **101**, 1449–1464.

7. Huffer, F. and McKeague, I. W. (1991). Weighted least squares estimation for Aalen's additive risk model, *Journal of the American Statistical Association*, **86**, 114–129.

8. McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model, *Biometrika*, **81**, 501–514.

9. Martinussen, T. and Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data*, Springer-Verlag, New York.