

The Interval-Censored Semi-Nonparametric Mixed Proportional Hazard Model and its Implementation in EasyReg

Herman J. Bierens

May 10, 2005

1 The mixed proportional hazard model

This *EasyReg* module (SNPSURVIVAL) estimates a semi-nonparametric (SNP) mixed proportional hazard model for a duration T , conditional on a vector of covariates, according to the approach in Bierens (2005).

Given a vector $X \in \mathbb{R}^k$ of covariates, and a non-negative independent random variable V representing unobserved heterogeneity, the conditional hazard function is defined as

$$\tilde{\lambda}(t|X, V) = \lim_{\delta \downarrow 0} \frac{P[T \in [t, t + \delta) \mid X, V, T \geq t]}{\delta}.$$

Then

$$P[T > t|X, V] = \exp\left(-\int_0^t \tilde{\lambda}(\tau|X, V) d\tau\right).$$

In the case of a mixed proportional hazard model it is assumed that $\tilde{\lambda}(t|X, V)$ takes the form

$$\tilde{\lambda}(t|X) = V \exp(\beta' X) \lambda(t|\alpha),$$

where $\exp(\beta' X)$ is the systematic hazard and $\lambda(t|\alpha)$ is the baseline hazard function depending on a parameter (vector) α . Then conditional on X alone,

$$P[T > t|X] = \int_0^\infty \exp(-v \exp(\beta' X) \Lambda(t|\alpha)) dG(v),$$

where $G(v)$ is the **unknown** distribution function of the unobserved heterogeneity V , and

$$\Lambda(t|\alpha) = \int_0^t \lambda(\tau|\alpha) d\tau$$

is the integrated hazard. The conditional probability $P[T > t|X]$ is known as the conditional survival function.

The conditional survival function involved can be written as

$$S(t|\alpha, \beta'X, h) = H(\exp(-\exp(\beta'X)\Lambda(t|\alpha))),$$

where

$$H(u) = \int_0^\infty u^v dG(v), \quad u \in [0, 1],$$

is a distribution function on $[0, 1]$, with density function

$$h(u) = H'(u) = \int_0^\infty v u^{v-1} dG(v), \quad u \in (0, 1].$$

The integrated baseline hazard $\Lambda(t|\alpha)$ will be specified such that there is no need to include a constant in X .

2 Interval censoring

It is assumed that the duration T is not observed directly, but only in the form of M dummy variables corresponding to intervals:

$$D_i = I(T \in (b_{i-1}, b_i]), \quad i = 1, \dots, M,$$

say, where $I(\cdot)$ is the indicator function, $I(true) = 1$, $I(false) = 0$, and $0 = b_0 < b_1 < \dots < b_M$ (called the bracket points). The intervals $(b_{i-1}, b_i]$ may be replaced by $[b_{i-1}, b_i)$. However, these intervals should be disjoint. If the duration T is right-censored as well, the bracket point b_M should be such that, for all observations, T is not right-censored if $T \in (0, b_M]$.

It is recommended to indicate the bracket values in the name of D_i , for example, use names like "Dummy T in (2,4]" or "I(T in (2,4])". *EasyReg* will then extract the lower and upper bounds of the bracket from the name.

Note that for $i = 1, 2, \dots, M$,

$$\begin{aligned} P[D_i = 1|X] &= S(b_{i-1}|\alpha, \beta'X, h) - S(b_i|\alpha, \beta'X, h) \\ &= H(\exp(-\exp(\beta'X)\Lambda(b_{i-1}|\alpha))) - H(\exp(-\exp(\beta'X)\Lambda(b_i|\alpha))), \end{aligned}$$

and

$$P \left[\sum_{i=1}^M D_i = 0 \middle| X \right] = S(b_M | \beta' X, h) = H(\exp(-\exp(\beta' X) \Lambda(b_M | \alpha))).$$

3 Hazard function options

3.1 Piecewise linear integrated baseline hazard

Because these probabilities depend on the values of $\Lambda(b_i | \alpha)$ in M bracket points b_i only, we may without loss of generality parametrize $\Lambda(t | \alpha)$ as a piecewise linear function:

$$\begin{aligned} \Lambda(t | \alpha) &= \Lambda(b_{i-1} | \alpha) + \alpha_i (t - b_{i-1}) & (1) \\ &= \sum_{k=1}^{i-1} \alpha_k (b_k - b_{k-1}) + \alpha_i (t - b_{i-1}) \text{ for } t \in (b_{i-1}, b_i], \\ \alpha_i &> 0 \text{ for } i = 1, \dots, M, \alpha = (\alpha_1, \dots, \alpha_M)' \in \mathbb{R}^M. \end{aligned}$$

There are other equivalent ways to specify $\Lambda(b_i | \alpha)$, but the advantage of the specification (1) is that the null hypothesis $\alpha_1 = \dots = \alpha_M$ corresponds to the integrated Weibull hazard $\Lambda(t | \alpha) = \alpha_1 t$. In that case $\exp(\beta' X) \Lambda(t | \alpha) = \exp(\ln(\alpha_1) + \beta' X) t$, so that $\ln(\alpha_1)$ acts as a constant term in the systematic hazard.

The specification (1) is adopted in *EasyReg* as the default option, and is called the piecewise linear integrated baseline hazard.

Next to this specification, you have four other options for the baseline hazard:

3.2 Weibull baseline hazard

$$\begin{aligned} \lambda(t | \alpha) &= \alpha_1 \alpha_2 t^{\alpha_2 - 1}, & (2) \\ \alpha_1 &> 0, \alpha_2 > 0, \alpha = (\alpha_1, \alpha_2)', \\ \Lambda(t | \alpha) &= \int_0^t \lambda(\tau | \alpha) d\tau = \alpha_1 t^{\alpha_2}. \end{aligned}$$

The reason for the scale factor α_1 here and below is that X does not contain a constant, hence $\ln(\alpha_1)$ plays the role of constant: $\exp(\beta' X) \alpha_1 t^{\alpha_2} =$

$\exp(\ln(\alpha_1) + \beta'X)t^{\alpha_2}$. Note that if $\alpha_2 > 1$ then $\lambda(t|\alpha)$ is monotonic increasing, starting from $\lambda(0|\alpha) = 0$, and if $\alpha_2 < 1$ then $\lambda(t|\alpha)$ is monotonic decreasing, and $\lambda(0|\alpha) = \infty$.

3.3 Generalized Weibull baseline hazard

$$\begin{aligned}\lambda(t|\alpha) &= \alpha_1 \alpha_2 (\alpha_3 + t)^{\alpha_2 - 1}, \\ \alpha_1 > 0, \alpha_2 > 0, \alpha_3 > 0, \alpha &= (\alpha_1, \alpha_2, \alpha_3)' \\ \Lambda(t|\alpha) &= \int_0^t \lambda(\tau|\alpha) d\tau = \alpha_1 ((\alpha_3 + t)^{\alpha_2} - \alpha_3^{\alpha_2}).\end{aligned}\tag{3}$$

The reason for the extra parameter α_3 is to allow $0 < \lambda(0|\alpha) < \infty$.

3.4 Unimodal baseline hazard

$$\begin{aligned}\lambda(t|\alpha) &= \frac{2\alpha_1 t}{\alpha_2^2 + t^2}, \alpha_2 = \arg \max_{t \geq 0} \lambda(t|\alpha), \\ \alpha_1 > 0, \alpha_2 > 0, \alpha &= (\alpha_1, \alpha_2)' \\ \Lambda(t|\alpha) &= \int_0^t \lambda(\tau|\alpha) d\tau = \alpha_1 \cdot \ln \left(\frac{\alpha_2^2 + t^2}{\alpha_2^2} \right).\end{aligned}\tag{4}$$

This hazard function is increasing on $(0, \alpha_2)$ and decreasing on (α_2, ∞) , with $\lambda(0|\alpha) = \lambda(\infty|\alpha) = 0$.

3.5 Generalized unimodal baseline hazard

$$\begin{aligned}\lambda(t|\alpha) &= \frac{2\alpha_1 (\alpha_3 + t)}{(\alpha_2 + \alpha_3)^2 + (\alpha_3 + t)^2}, \alpha_2 = \arg \max_{t \geq 0} \lambda(t|\alpha), \\ \alpha_1 > 0, \alpha_2 > 0, \alpha_3 > 0, \alpha &= (\alpha_1, \alpha_2, \alpha_3)' \\ \Lambda(t|\alpha) &= \int_0^t \lambda(\tau|\alpha) d\tau = \alpha_1 \cdot \ln \left(\frac{(\alpha_2 + \alpha_3)^2 + (\alpha_3 + t)^2}{(\alpha_2 + \alpha_3)^2 + \alpha_3^2} \right).\end{aligned}\tag{5}$$

Again, the reason for the extra parameter α_3 is to allow $\lambda(0|\alpha) > 0$.

4 SNP approximation of the density $h(u)$

The unknown density $h(u)$ and its distribution function $H(u)$ are parametrized as

$$h_q(u|\delta) = \frac{(1 + \sum_{n=1}^q \delta_n \rho_n(u))^2}{1 + \sum_{n=1}^q \delta_n^2}, \quad \delta = (\delta_1, \dots, \delta_q)',$$

$$H_q(u|\delta) = \int_0^u h_q(v|\delta) dv,$$

respectively, where the $\rho_n(u)$'s are orthonormal Legendre polynomials of order n on $[0, 1]$:

$$\int_0^1 \rho_k(u) \rho_m(u) du = \begin{cases} 0 & \text{if } k \neq m, \\ 1 & \text{if } k = m. \end{cases}$$

For $n \geq 2$ the Legendre polynomials can be computed recursively by

$$\rho_n(u) = \frac{\sqrt{2n-1}\sqrt{2n+1}}{n}(2u-1)\rho_{n-1}(u) - \frac{(n-1)\sqrt{2n+1}}{n\sqrt{2n-3}}\rho_{n-2}(u),$$

starting from $\rho_0(u) = 1$, $\rho_1(u) = \sqrt{3}(2u-1)$.

In order for α , β and δ to be identified we need to fix the location and scale of $H_q(u|\delta)$, by imposing the following moment restrictions on $h_q(u|\delta)$:

$$\int_0^1 \ln(\ln(1/u)) h_q(u|\delta) du = \int_0^1 \ln(\ln(1/u)) du,$$

$$\int_0^1 (\ln(\ln(1/u)))^2 h_q(u|\delta) du = \int_0^1 (\ln(\ln(1/u)))^2 du.$$

These moment restrictions are imposed in *EasyReg* by penalizing the log-likelihood function with the following two penalty terms:

$$\Pi_{q,N}^{(1)}(\delta) = -N(1 + \delta'\delta)^4 \left(\int_0^1 \ln(\ln(1/u)) h_q(u|\delta) du - \int_0^1 \ln(\ln(1/u)) du \right)^4,$$

$$\Pi_{q,N}^{(2)}(\delta) = -N(1 + \delta'\delta)^4 \left(\int_0^1 (\ln(\ln(1/u)))^2 h_q(u|\delta) du - \int_0^1 (\ln(\ln(1/u)))^2 du \right)^4,$$

where N is the sample size.

5 Three-step ML estimation

EasyReg estimates the parameter vectors α , β and δ in three steps. In first instance the parameters α_i are fixed to $\alpha_i = 1$, except that in the cases (4) and (5) $\alpha_2 = b_1$, and δ is set equal to a zero vector

The (quasi-)maximum likelihood estimator $\tilde{\beta}_0$ of β in the first step will be used as starting values in the second step, together with the initial values of α_i , keeping $\delta = 0$. This step yields (quasi-)maximum likelihood estimators $\tilde{\alpha}_1$ of α and $\tilde{\beta}_1$ of β .

Again, these estimates are merely used as starting values in the final step, where $h(u)$ and $H(u)$ are approximated by $h_q(u|\delta)$ and $H_q(u|\delta)$, respectively

If you check "Batch mode" these three rounds are conducted automatically, where in each round the iteration is automatically restarted until the log-likelihood does not change anymore. This option is recommended for big jobs.

If $q \rightarrow \infty$ with the sample size N then under some regularity conditions the final ML estimators $\hat{\alpha}$, $\hat{\beta}$ are weakly consistent: $p \lim_{N \rightarrow \infty} \hat{\alpha} = \alpha$ and $p \lim_{N \rightarrow \infty} \hat{\beta} = \beta$. However, the asymptotic normality results for $(\hat{\beta}, \hat{\alpha}, \hat{\delta})$ given by *EasyReg* are only valid if for a fixed q and a $\delta_0 \in \mathbb{R}^q$, $h_q(u|\delta_0) \equiv h(u)$.

6 Reference

Bierens, H. J. (2005), "Semi-Nonparametric Modeling of Densities on the Unit Interval, with Applications to Censored Mixed Proportional Hazard Models and Ordered Probability Models: Identification and Consistency Results", working paper, Department of Economics, PennState University.