

EasyReg Module SNPSURVIVAL1

1 The mixed proportional hazard model

Let T be a duration, and let X be a vector of covariates. As is well-known, the conditional hazard function is defined as

$$\frac{f(t|X)}{1 - F(t|X)} = \lambda(t, X),$$

where $F(t|X) = P[T \leq t|X]$, $f(t|X)$ is the corresponding conditional density function, and $\int_0^\infty \lambda(\tau, X)d\tau = \infty$. Then the conditional survival function is

$$S(t|X) = 1 - F(t|X) = \exp\left(-\int_0^t \lambda(\tau, X)d\tau\right).$$

The mixed proportional hazard model assumes that the conditional survival function takes the form

$$\begin{aligned} S(t|X, \alpha, \beta) &= S(t|X) \\ &= E\left[\exp\left(-\exp(\beta'X + U)\int_0^t \lambda(\tau|\alpha)d\tau\right)\middle|X\right], \end{aligned} \tag{1}$$

where U represents unobserved heterogeneity, which is independent of X , $\lambda(t|\alpha)$ is the baseline hazard function depending on a parameter (vector) α , and $\exp(\beta'X)$ is the systematic hazard function. Denoting the distribution function of $V = \exp(U)$ by $G(v)$, and the integrated baseline hazard by

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

we have

$$\begin{aligned} S(t|X, \alpha, \beta, h) &= P[T > t|X, \alpha, \beta, h] \\ &= \int_0^\infty \exp(-v \cdot \exp(\beta'X)\Lambda(t|\alpha)) dG(v) \\ &= \int_0^\infty (\exp(-\exp(\beta'X)\Lambda(t|\alpha)))^v dG(v) \\ &= H(\exp(-\exp(\beta'X)\Lambda(t|\alpha))), \end{aligned} \tag{2}$$

where

$$H(u) = \int_0^\infty u^v dG(v), \quad u \in [0, 1], \quad (3)$$

is a distribution function on $[0, 1]$. Note that the intergrated baseline hazard $\Lambda(t|\alpha)$ has to be monotonic, and $\Lambda(\infty|\alpha) = \infty$, as otherwise (2) is not a valid probability for all $t > 0$.

If the unobserved heterogeneity variable V satisfies $E[V] < \infty$ then for $u \in (0, 1]$,

$$\int_0^\infty vu^{v-1} dG(v) \leq u^{-1} \int_0^\infty v dG(v) < \infty, \quad (4)$$

so that by the mean value and dominated convergence theorems, $H(u)$ is differentiable on $(0, 1)$, with density function

$$h(u) = \int_0^\infty vu^{v-1} dG(v). \quad (5)$$

This is the reason for the argument h in the left-hand side of (2). Moreover, (4) implies that $h(u)$ is finite and continuous¹ on $(0, 1]$. Furthermore, note that absence of unobserved heterogeneity, i.e., $P[V = 1] = 1$, is equivalent to the case $h(u) \equiv 1$.

Let the true conditional survival function be

$$\begin{aligned} S(t|X, \alpha_0, \beta_0, h_0) &= \int_0^\infty \exp(-v \cdot \exp(\beta_0' X) \Lambda(t|\alpha_0)) dG_0(v) \\ &= H_0(\exp(-\exp(\beta_0' X) \Lambda(t|\alpha_0))) \end{aligned} \quad (6)$$

where $H_0(u) = \int_0^u h_0(v) dv = \int_0^\infty u^v dG_0(v)$. In the expressions (2) and (6), h and h_0 should be interpreted as unknown parameters contained in a parameter space of density functions on $(0, 1]$.

2 Censoring

Usually the duration T is only observed up to an upper bound \bar{T} , which may vary per individual. This is called (right) censoring. Indicate censoring by the dummy variable

$$C = I(T > \bar{T}),$$

¹The continuity also follows from the dominated convergence theorem.

where $I(\cdot)$ is the indicator function, i.e., $I(true) = 1$, $I(false) = 0$. Assuming that \bar{T} is exogeneous and does not depend on the covariates, it follows from (6) that

$$\begin{aligned} P[C = 1|X] &= S(\bar{T}|X, \alpha_0, \beta_0, h_0) \\ &= H_0(\exp(-\exp(\beta'_0 X)\Lambda(\bar{T}|\alpha_0))), \end{aligned} \quad (7)$$

and

$$\begin{aligned} P[T \leq t|X, C = 0] &= \frac{P[T \leq t, T \leq \bar{T}|X]}{P[C = 0|X]} = \frac{P[T \leq \min(t, \bar{T})|X]}{P[C = 0|X]} \\ &= \frac{1 - H_0(\exp(-\exp(\beta'_0 X)\Lambda(\min(t, \bar{T})|\alpha_0)))}{1 - H_0(\exp(-\exp(\beta'_0 X)\Lambda(\bar{T}|\alpha_0)))}. \end{aligned}$$

Taking the derivative to t yields the density $f(t|X, C = 0, \alpha_0, \beta_0, h_0)$ of T conditional on the covariates and absence of censoring:

$$\begin{aligned} &f(t|C = 0, X, \alpha_0, \beta_0, h_0) \\ &= \frac{h_0(\exp(-\exp(\beta'_0 X)\Lambda(t|\alpha_0))) \exp(-\exp(\beta'_0 X)\Lambda(t|\alpha_0))}{1 - H_0(\exp(-\exp(\beta'_0 X)\Lambda(\bar{T}|\alpha_0)))} \\ &\quad \times \exp(\beta'_0 X)\lambda(t|\alpha_0) \text{ if } t \leq \bar{T}, \\ &= 0 \text{ if } t > \bar{T}. \end{aligned} \quad (8)$$

3 Nonparametric identification

Elbers and Ridder (1982) have shown that if X does not contain a constant,

$$\Lambda(t|\alpha) = \Lambda(t|\alpha_0) \text{ for all } t > 0 \text{ implies } \alpha = \alpha_0, \quad (9)$$

and

$$\int_0^\infty v dG_0(v) = \int_0^\infty v dG(v) = 1 \quad (10)$$

(which by (5) is equivalent to confining the parameter space for the densities h to a space of densities h on $(0, 1]$ satisfying $h(1) = 1$), then the MPH model is nonparametrically identified, in the sense that

$$S(T|X, \alpha, \beta_0, h) = S(T|X, \alpha_0, \beta_0, h_0) \text{ a.s.}$$

implies $\alpha = \alpha_0$ and $G = G_0$, hence $h(u) = h_0(u)$ a.e. on $[0, 1]$. Heckman and Singer (1984) provide an alternative identification proof, and propose to parametrize G_0 as a discrete distribution: $G_0(v) = \sum_{i=1}^q I(v \leq \theta_i) p_i$, with $I(\cdot)$ the indicator function, where $\theta_i > 0$, $p_i > 0$, and $\sum_{i=1}^q p_i = 1$. Thus, they implicitly specify $h_0(u) = \sum_{i=1}^q \theta_i u^{\theta_i - 1} p_i$.

The nonparametric identification of the MPH model hinges on the assumption that T is observed directly if T is not right-censored.

4 The log-likelihood function

Given i.i.d. observations $\{T_j, D_j, C_j\}_{j=1}^N$ on (T, D, C) , and assuming that $T_j = \bar{T}_j$ if $C_j = 1$, the log-likelihood function is

$$\begin{aligned} & \ln(L_N(\alpha, \beta, h)) \\ &= \sum_{j=1}^N C_j \ln(H(\exp(-\exp(\beta' X_j) \Lambda(T_j|\alpha)))) \\ &+ \sum_{j=1}^N (1 - C_j) \ln(h(\exp(-\exp(\beta' X_j) \Lambda(T_j|\alpha)))) \\ &- \sum_{j=1}^N (1 - C_j) \exp(\beta' X_j) \Lambda(T_j|\alpha) \\ &+ \sum_{j=1}^N (1 - C_j) \beta' X_j + \sum_{j=1}^N (1 - C_j) \ln(\lambda(T_j|\alpha)) \end{aligned}$$

At this point the density $h(u)$ is treated as a parameter.

5 Specifying the unobserved heterogeneity distribution

5.1 Flexible functional form

In the case of right-censoring it may not be true that $h(u) = h_0(u)$ a.e. on $(0, 1]$. This is not too serious a problem, though, because we will model h_0 in

a flexible way, but involving only a finite number of parameters, similar to the approach in Bierens (2007).

In particular, we will assume that $h_0(u)$ belongs to the space of density functions of the type

$$h_q(u) = h_q(u|\delta) = \frac{(1 + \sum_{k=1}^q \delta_k \rho_k(u))^2}{1 + \sum_{k=1}^q \delta_k^2}, \quad \delta = (\delta_1, \dots, \delta_q)', \quad (11)$$

for a given **fixed** but unknown natural number q , where the $\rho_k(u)$'s are orthonormal Legendre polynomials on the unit interval:

$$\int_0^1 \rho_k(u) \rho_m(u) du = \begin{cases} 0 & \text{if } k \neq m, \\ 1 & \text{if } k = m. \end{cases} \quad (12)$$

Thus, it will be assumed that $h_0(u) \equiv h_q(u|\delta_0)$ for some unique $\delta_0 \in \mathbb{R}^q$.

The Legendre polynomials can easily be generated recursively by

$$\rho_n(u) = \frac{\sqrt{2n-1}\sqrt{2n+1}}{n}(2u-1)\rho_{n-1}(u) - \frac{(n-1)\sqrt{2n+1}}{n\sqrt{2n-3}}\rho_{n-2}(u), \quad (13)$$

for $n \geq 2$, starting from

$$\rho_0(u) = 1, \quad \rho_1(u) = \sqrt{3}(2u-1). \quad (14)$$

The identification condition $h_q(1|\delta) = 1$ can be imposed by restricting δ_1 to be

$$\begin{aligned} \delta_1 &= \frac{1}{2} \sqrt{2 \left(1 + \sum_{k=2}^q \delta_k^2 \right) + \left(1 + \sum_{k=2}^q \delta_k \rho_k(1) \right)^2} \\ &\quad - \frac{\sqrt{3}}{2} \left(1 + \sum_{k=2}^q \delta_k \rho_k(1) \right) \\ &= \frac{1}{2} \sqrt{2 \left(1 + \sum_{k=2}^q \delta_k^2 \right) + \left(1 + \sum_{k=2}^q \delta_k \sqrt{2k+1} \right)^2} \\ &\quad - \frac{\sqrt{3}}{2} \left(1 + \sum_{k=2}^q \delta_k \sqrt{2k+1} \right), \end{aligned} \quad (15)$$

where the latter equality follows from the fact that $\rho_k(1) = \sqrt{2k+1}$. See Bierens (2007) for further details and the motivation for using densities of the type (11).

5.2 The Gamma distribution

A popular parametric specification of the distribution of V is the Gamma(δ, τ) distribution, because the Laplace transform of the Gamma(δ, τ) distribution $G(v)$ takes a closed form:

$$\mathcal{L}(s) = E[\exp(-s.V)] = \int_0^\infty \exp(-s.v)dG(v) = (1 + \tau.s)^{-\delta}.$$

The Laplace transform $\mathcal{L}(s)$ is related to the distribution function $H(u) = \int_0^\infty u^v dG(v)$ by the equality

$$\mathcal{L}(s) = H(\exp(-s)) = (1 + \tau.s)^{-\delta}.$$

Hence

$$\begin{aligned} H(u) &= (1 + \tau.\ln(1/u))^{-\delta} \\ h(u) &= H'(u) = \frac{\tau.\delta}{u} \left(\frac{1}{1 + \tau.\ln(1/u)} \right)^{\delta+1} = \frac{\tau.\delta}{u} H(u)^{(\delta+1)/\delta}. \end{aligned}$$

Because of the presence of a scale factor in the baseline hazard (see below), the parameter τ has to be fixed to a constant, or made dependent on δ . To facilitate the comparison with the previous flexible specification, it will be assumed that $\tau = 1/\delta$, so that in this case $h(1) = 1$ as well. Thus

$$H(u|\delta) = (1 + \delta^{-1}.\ln(1/u))^{-\delta}, \quad h(u|\delta) = \frac{1}{u} H(u|\delta)^{(\delta+1)/\delta}. \quad (16)$$

6 Implementation in EasyReg

EasyReg module SNPSURVIVAL1 will ask you to select T , the right-censoring dummy variable C , and the covariates X . Moreover, for the density $h(u) = \int_0^\infty v u^{v-1} dG(v)$ you can choose either (11) or (16).

The following options for the baseline and integrated hazard of T are available.

6.1 Weibull hazard

$$\lambda(t|\alpha) = \alpha_1 \alpha_2 t^{\alpha_2 - 1}, \quad (17)$$

$$\alpha_1 > 0, \alpha_2 > 0, \alpha = (\alpha_1, \alpha_2)'$$

Integrated hazard:

$$\Lambda(t|\alpha) = \int_0^t \lambda(\tau|\alpha) d\tau = \alpha_1 t^{\alpha_2}.$$

6.2 Generalized Weibull hazard

If in the Weibull case $\alpha_2 < 1$ then $\lambda(0|\alpha) = \infty$, whereas if $\alpha_2 > 1$ then $\lambda(0|\alpha) = 0$. This may be too restrictive. The following generalized Weibull hazard specification satisfies $0 < \lambda(0|\alpha) < \infty$:

$$\lambda(t|\alpha) = \alpha_1 \alpha_2 (\alpha_3 + t)^{\alpha_2 - 1},$$

$$\alpha_1 > 0, \alpha_2 > 0, \alpha_3 > 0, \alpha = (\alpha_1, \alpha_2, \alpha_3)'$$

Integrated hazard:

$$\Lambda(t|\alpha) = \int_0^t \lambda(\tau|\alpha) d\tau = \alpha_1 ((\alpha_3 + t)^{\alpha_2} - \alpha_3^{\alpha_2}).$$

6.3 Unimodal hazard

$$\lambda(t|\alpha) = \frac{2\alpha_1 t}{\alpha_2^2 + t^2}, \alpha_2 = \arg \max_{t \geq 0} \lambda(t|\alpha),$$

$$\alpha_1 > 0, \alpha_2 > 0, \alpha = (\alpha_1, \alpha_2)'$$

Integrated hazard:

$$\Lambda(t|\alpha) = \int_0^t \lambda(\tau|\alpha) d\tau = \alpha_1 \cdot \ln \left(\frac{\alpha_2^2 + t^2}{\alpha_2^2} \right).$$

6.4 Generalized unimodal hazard

In the unimodal hazard case, $\lambda(0|\alpha) = 0$. Again, this may be too restrictive. The following generalized unimodal hazard specification allows $\lambda(0|\alpha) > 0$:

$$\lambda(t|\alpha) = \frac{2\alpha_1 (\alpha_3 + t)}{(\alpha_2 + \alpha_3)^2 + (\alpha_3 + t)^2}, \alpha_2 = \arg \max_{t \geq 0} \lambda(t|\alpha),$$

$$\alpha_1 > 0, \alpha_2 > 0, \alpha_3 > 0, \alpha = (\alpha_1, \alpha_2, \alpha_3)'$$

Integrated hazard:

$$\Lambda(t|\alpha) = \int_0^t \lambda(\tau|\alpha) d\tau = \alpha_1 \cdot \ln \left(\frac{(\alpha_2 + \alpha_3)^2 + (\alpha_3 + t)^2}{(\alpha_2 + \alpha_3)^2 + \alpha_3^2} \right).$$

In all four cases the parameter α_1 acts as a scale factor. Therefore, $\ln(\alpha_1)$ acts as a constant term. Consequently, the vector X of covariates should **not** contain a constant.

7 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 $\alpha_i = 1$, 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.

References

- Bierens, H.J. (2007), "Semi-Nonparametric Interval Censored Mixed Proportional Hazard Models: Identification and Consistency Results", forthcoming in *Econometric Theory*.
- Elbers, C., and G. Ridder (1982), " True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model", *Review of Economic Studies*, 49, 403-409.
- Heckman, J. J., and B. Singer (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data", *Econometrica*, 52, 271-320.