

Modèles de durée - Approche non paramétrique

Contexte et objectifs.

$x_1, \dots, x_n$ sont des durées de vie, $c_1, \dots, c_n$ des instants de censure à droite, $t_1, \dots, t_n$ les observations : $t_i = \min\{x_i, c_i\}$ et $\delta_1, \dots, \delta_n$ les indicatrices de non censure : $\delta_i = \mathbb{1}_{\{x_i \leq c_i\}}$.

On suppose $x_1, \dots, x_n$ distribuées selon un modèle de durée de vie continue ; on souhaite estimer la fonction de survie $S(t)$ et la fonction de hasard cumulé $\Lambda(t)$ de façon non paramétrique.

On supposera en plus : $x_1, \dots, x_n$ sont indépendantes ; pour tout i , x_i et c_i sont indépendantes ; $t_1, \dots, t_n$ sont indépendantes.

Estimateurs pour la fonction de survie & le hasard cumulé.

o L'estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier de $S(t)$ est : $\hat{S}_{KM}(t) = \prod_{i:t_{(i)} \leq t} (1 - d_i/r_i)$ où $t_{(i)}$ désigne la statistique d'ordre de rang i de la série $t_1, \dots, t_n$ dont les répétitions ont été enlevées, d_i le nombre de décès en $t_{(i)}$ et r_i l'exposition à $t_{(i)}$ c'est à dire le nombre de survivants juste avant $t_{(i)}$.

★ La fonction $t \mapsto \hat{S}_{KM}(t)$ est constante par morceaux, continue à droite et ses discontinuités sont aux instants de décès.

★ L'estimateur de Kaplan-Meier est biaisé ; sa variance est souvent estimée par (formule de Greenwood) : $\hat{V}(\hat{S}_{KM}(t)) = \hat{S}_{KM}^2(t) \sum_{i:t_{(i)} \leq t} d_i/[r_i(r_i - d_i)]$.

★ Si S est continue en $t \in [0; \tau]$ et $\lim_{\tau^-} S > 0$, quand n est grand, $\hat{S}_{KM}(t)$ est approximativement distribué selon une loi normale de moyenne $S(t)$ et de variance $\hat{V}(\hat{S}_{KM}(t))$. Un intervalle de confiance de $S(t)$ au niveau $1 - \alpha$ est alors : $IC_{1-\alpha}(S(t)) = \hat{S}_{KM}(t) \pm z_{\alpha/2} \sqrt{\hat{V}(\hat{S}_{KM}(t))}$.

o L'estimateur de Breslow

Les fonctions de hasard cumulé et de survie sont liées par : $\Lambda(t) = -\log S(t)$; l'estimateur de Kaplan-Meier de $S(t)$ induit donc un estimateur non paramétrique de $\Lambda(t)$ appelé l'estimateur de Breslow : $\hat{\Lambda}_B(t) = -\log \hat{S}_{KM}(t)$.

★ Un estimateur de la variance de $\hat{\Lambda}_B(t)$ est $\hat{V}(\hat{\Lambda}_B(t)) = \sum_{i:t_{(i)} \leq t} d_i/[r_i(r_i - d_i)]$.

o La méthode actuarielle

Pour $\tau_0 = 0 < \tau_1 < \dots < \tau_{k-1} < \tau_k = +\infty$ on note d_i le nombre de décès, ν_i le nombre de sujets censurés, r_i le nombre de sujets à risque dans l'intervalle $[\tau_{i-1}, \tau_i[$ et n_i le nombre de sujets vivants à τ_i .

En supposant que dans chaque intervalle les censures sont réparties de façon uniforme, chaque individu censuré de $[\tau_{i-1}, \tau_i[$ compte pour moitié et (hypothèse actuarielle) : $r_i = n_{i-1} - \nu_i/2$.

L'estimateur actuariel de $S(t)$ est alors : $\hat{S}_a(t) = \prod_{i:\tau_i \leq t} (1 - d_i/r_i)$; la variance de cet estimateur peut être estimée par : $\hat{V}(\hat{S}_a(t)) = \hat{S}_a^2(t) \sum_{i:\tau_i \leq t} d_i/[r_i(r_i - d_i)]$.

o L'estimateur de Nelson-Aalen

L'estimateur de Nelson-Aalen de $\Lambda(t)$ est : $\hat{\Lambda}_{NA}(t) = \sum_{i:t_{(i)} \leq t} d_i/r_i$.

★ A taille d'échantillon finie $\hat{\Lambda}_{NA}(t)$ sous estime en moyenne $\Lambda(t)$; mais sous certaines conditions l'estimateur de Nelson-Aalen est asymptotiquement sans biais et asymptotiquement normal.

★ Un estimateur de la variance de $\hat{\Lambda}_{NA}(t)$ est : $\hat{V}(\hat{\Lambda}_{NA}(t)) = \sum_{i:t_{(i)} \leq t} d_i/r_i^2$.

★ En tout point l'estimateur de Nelson-Aalen est inférieur à celui de Breslow : $\hat{\Lambda}_{NA}(t) < \hat{\Lambda}_B(t)$.

o L'estimateur d'Harrington-Fleming

La fonction de survie et le hasard cumulé sont liés par : $S(t) = \exp\{-\Lambda(t)\}$; l'estimateur de Nelson-Aalen du hasard cumulé induit donc un estimateur non paramétrique de la survie : $\hat{S}_{HF}(t) = \exp\{-\hat{\Lambda}_{NA}(t)\}$ appelé l'estimateur d'Harrington-Fleming.

★ Un estimateur de la variance de $\hat{S}_{HF}(t)$ est : $\hat{V}(\hat{S}_{HF}(t)) = \exp\{-2 \sum_{i:t_{(i)} \leq t} d_i/r_i\} \times \left(\sum_{i:t_{(i)} \leq t} d_i/r_i^2 \right)$.

Comparaison des fonctions de survie.

o Le test de Mantel-Haenszel

On suppose les observations issues de deux groupes, on note S_j la fonction de survie dans le groupe j ($j \in \{1, 2\}$) et on souhaite tester $\mathcal{H}_0 : S_1 = S_2$ contre $\mathcal{H}_1 : S_1 \neq S_2$.

Pour un instant $t_{(i)}$ notons : $d_{i,j}$ le nombre de décès dans le groupe j , $r_{i,j}$ l'exposition dans le groupe j , $d_i = d_{i,1} + d_{i,2}$ le nombre total de décès et $r_i = r_{i,1} + r_{i,2}$ l'exposition totale.

En notant : $D_j = \sum_i d_{i,j}$ et $E_j = \sum_i d_i r_{i,j} / r_i$, sous $\mathcal{H}_0$, la statistique $X = \sum_{j=1}^2 (D_j - E_j)^2 / E_j$ est approximativement distribuée selon χ_1^2 ; on rejette donc $\mathcal{H}_0$ au seuil α lorsque X est supérieure à $\chi_{1;1-\alpha}^2$.
