
Travaux dirigés : modèles de durée
Séance n°2 du 30 Novembre 2015 - Corrigé

Exercice 1 Manipulation de l'estimateur de Kaplan-Meier.

Soit un échantillon de n individus i.i.d. de durée de vie $T_1, \dots, T_n$. Chaque observation i est soumise à censure à droite C_i , supposée indépendante et non-informative. On observe (Y_i, D_i) pour chaque i , où $Y_i = T_i \wedge C_i$ et $D_i = \mathbb{1}_{\{T_i \leq C_i\}}$. On note S la fonction de survie de T et G celle de C .

1. En notant $H_1(t) = \mathbb{P}(Y > t, D = 1)$ et $H(t) = \mathbb{P}(Y > t)$, montrer que

$$\frac{dS(t)}{S(t-)} = \frac{dH_1(t)}{H(t-)}.$$

2. Retrouver à partir de cette expression, l'estimateur non-paramétrique de Nelson-Aalen pour la fonction de hasard cumulée $\Lambda(t)$.
3. Rappeler l'expression de la fonction de hasard $h(t)$, de la fonction de hasard cumulée $\Lambda(t)$ et de la fonction de survie d'une loi discrète. En déduire l'expression de l'estimateur de Kaplan-Meier de la fonction de survie.
4. En suivant un raisonnement similaire avec $H_1(t) = \mathbb{P}(Y > t, D = 1 | Y > L)$ et $H(t) = \mathbb{P}(L \leq t < Y | Y > L)$, donner l'expression de l'estimateur de Kaplan-Meier avec troncature à gauche L indépendante (T indépendant de (L, C) et C indépendant de L).
5. En utilisant directement l'expression de la première question, exprimer l'estimateur de Kaplan-Meier sous la forme d'une somme faisant intervenir l'estimateur de G .

Réponse de l'exercice 1.

1. On a

$$\begin{aligned} H(t) &= \mathbb{P}(Y > t) = \mathbb{P}(T > t, C > t) \\ &= S(t) G(t), \end{aligned}$$

et comme $C \perp\!\!\!\perp T$

$$\begin{aligned} H_1(t) &= \mathbb{E}[\mathbb{1}_{\{T > t\}} \mathbb{1}_{\{C \geq T\}}] \\ &= \mathbb{E}[\mathbb{1}_{\{T > t\}} \mathbb{E}[\mathbb{1}_{\{C \geq T\}} | T]] \\ &= \mathbb{E}[\mathbb{1}_{\{T > t\}} G(T-)] \\ &= \int_t^\infty G(u-) dF(u). \end{aligned}$$

On écrit alors

$$dH_1(t) = -G(t-) dF(t) = G(t-) dS(t).$$

D'où le résultat.

2. Par définition,

$$\Lambda(t) = - \int_0^t \frac{dS(t)}{S(t-)} = - \int_0^t \frac{dH_1(t)}{H(t-)}.$$

Puisque les quantités Y_i et D_i sont observées, les estimateurs empiriques de $H_1(t)$ et $H(t)$ sont donnés par

$$\hat{H}_1(t) = \frac{1}{n} \sum_{i=1}^n D_i \mathbb{1}_{\{Y_i > t\}} \text{ et } \hat{H}(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{Y_i > t\}}.$$

Naturellement, l'estimateur de Nelson-Aalen apparaît en remplaçant $H_1(t)$ et $H(t)$ par leur estimateur, i.e.

$$\begin{aligned} \hat{\Lambda}(t) &= - \int_0^t \frac{d\hat{H}_1(t)}{\hat{H}(t-)} \\ &= \frac{1}{n} \sum_{i=1}^n \frac{D_i \mathbb{1}_{\{Y_i \leq t\}}}{\hat{H}(Y_i-)} \\ &= \sum_{i=1}^n \frac{D_i \mathbb{1}_{\{Y_i \leq t\}}}{\sum_{j=1}^n \mathbb{1}_{\{Y_j \geq Y_i\}}} \\ &= \sum_{Y_i \leq t}^n \frac{D_i}{\sum_{j=1}^n \mathbb{1}_{\{Y_j \geq Y_i\}}} \end{aligned}$$

3. En temps discret, la fonction de hasard est définie par

$$h(t) = \frac{\mathbb{P}(T = t)}{S(t)},$$

et la fonction de hasard cumulée est

$$\Lambda(t) = \sum_{s \leq t} h(s).$$

La fonction de survie s'écrit

$$S(t) = \prod_{s \leq t} (1 - h(s)) = \prod_{s \leq t} (1 - \Delta\Lambda(s)).$$

On retrouve ainsi l'expression de l'estimateur de Kaplan-Meier

$$\begin{aligned} \hat{S}(t) &= \prod_{Y_i \leq t} \left(1 - \Delta\hat{\Lambda}(T_i)\right) \\ &= \prod_{Y_i \leq t} \left(1 - \frac{D_i}{\sum_{j=1}^n \mathbb{1}_{\{Y_j \geq Y_i\}}}\right). \end{aligned}$$

4. On a immédiatement

$$\begin{aligned} H(t) &= \mathbb{P}(L \leq t < Y | Y > L) = \frac{\mathbb{P}(T > t, C > t, L \leq t)}{\mathbb{P}(Y > L)} \\ &= S(t) G(t) \frac{\mathbb{P}(L \leq t)}{\mathbb{P}(Y > L)}. \end{aligned}$$

En conditionnant par T et par indépendance avec L

$$\begin{aligned} H_1(t) &= \frac{\mathbb{E}[\mathbb{1}_{\{T > t\}} \mathbb{1}_{\{C \geq T\}} \mathbb{1}_{\{T > L\}}]}{\mathbb{P}(Y > L)} \\ &= \frac{\mathbb{E}[\mathbb{1}_{\{T > t\}} G(T-) \mathbb{P}(L \leq T)]}{\mathbb{P}(Y > L)} \end{aligned}$$

On écrit alors

$$dH_1(t) = \frac{1}{\mathbb{P}(Y > L)} G(t-) \mathbb{P}(L \leq t) dS(t).$$

Ainsi, on montre que

$$\frac{dH_1(t)}{H(t-)} = \frac{dS(t)}{S(t-)}.$$

En considérant les estimateurs empiriques de $H_1(t)$ et $H(t)$, on en déduit l'estimateur de Kaplan-Meier en présence de troncature à gauche tel que

$$\hat{S}(t) = \prod_{Y_i \leq t} \left(1 - \frac{D_i}{\sum_{j=1}^n \mathbb{1}_{\{L_i \leq Y_i \leq Y_j\}}} \right).$$

5. On écrit

$$dS(t) = \frac{dH_1(t)}{G(t-)}.$$

Ainsi en intégrant, il est possible d'exprimer l'estimateur de Kaplan-Meier sous la forme

$$\begin{aligned} \hat{S}(t) &= \frac{1}{n} \sum_{i=1}^n \frac{D_i}{\hat{G}(Y_i-)} \mathbb{1}_{\{Y_i > t\}} \\ &= \sum_{i=1}^n W_{i,n} \mathbb{1}_{\{Y_i > t\}}. \end{aligned}$$

On remarque que l'estimateur de Kaplan-Meier attribue des poids $W_{i,n}$ aux observations qui ne sont pas censurées. On peut montrer que ces poids en présence de censure s'écrivent, en ordonnant les valeurs de Y , $Y_{(1)} \leq \dots \leq Y_{(n)}$, et avec $D_{[i]}$ la valeur associée à $Y_{(i)}$

$$W_{[i],n} = \frac{D_{[i]}}{n - i + 1} \prod_{j=1}^{i-1} \frac{n - j}{n - j + 1}.$$

Cette expression sous forme de somme est commode car elle permet de fournir un estimateur pour les quantités du type $\Psi_\phi = \mathbb{E}[\phi(T)]$ avec ϕ une fonction intégrable (on parle d'intégrales Kaplan-Meier)

$$\hat{\Psi}_\phi = \sum_{i=1}^n W_{i,n} \phi(Y_i).$$

Exercice 2 Variance de l'estimateur de Kaplan-Meier.

Soit un échantillon de n individus i.i.d. de durée de vie $T_1, \dots, T_n$. Chaque observation i est soumise à censure à droite C_i , supposée indépendante et non-informative. On observe (Y_i, D_i) pour chaque i , où $Y_i = T_i \wedge C_i$ et $D_i = \mathbb{1}_{\{T_i \leq C_i\}}$. On note S la fonction de survie de T et G celle de C .

On rappelle que l'estimateur de Kaplan-Meier vérifie

$$\sqrt{n} \left(\hat{S} - S \right) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathbf{U} ,$$

avec $\mathbf{U}$ un processus gaussien centrée de variance-covariance

$$\rho(s, t) = -S(s) S(t) \int_0^{s \wedge t} \frac{dS(u)}{S(u)^2 G(u)}.$$

1. On note $H_1(t) = \mathbb{P}(Y > t, D = 1)$ et $H(t) = \mathbb{P}(Y > t)$. Puisque

$$\frac{dS(t)}{S(t-)} = \frac{dH_1(t)}{H(t-)} ,$$

donner un estimateur de la variance de $\hat{S}(t)$.

2. En notant $Y_{(i)}$ la statistique d'ordre i des Y et $D_{[i]}$ la valeur de D associée, réécrire cet estimateur de la variance et vérifier qu'il correspond à l'estimateur de Greenwood.
3. Proposer un intervalle de confiance de niveau α pour l'estimateur de Kaplan-Meier.
4. Déterminer la distribution asymptotique de $\hat{p}_t = \frac{\hat{S}(t+1)}{\hat{S}(t)}$, puis en déduire l'estimateur de sa variance (de type Greenwood).

Réponse de l'exercice 2.

1. La variance asymptotique de l'estimateur de Kaplan-Meier s'écrit

$$\begin{aligned} \mathbb{V}\text{ar} \left(\hat{S}(t) \right) &= -S(t)^2 \int_0^t \frac{dS(u)}{S(u)^2 G(u)} \\ &= -S(t)^2 \int_0^t \frac{dH_1(u)}{H(u) H(u-)} . \end{aligned}$$

En remplaçant $H_1(t)$ et $H(t)$ par leur estimateur empirique

$$\hat{H}_1(t) = \frac{1}{n} \sum_{i=1}^n D_i \mathbb{1}_{\{Y_i > t\}} \text{ et } \hat{H}(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{Y_i > t\}} ,$$

on obtient l'estimateur suivant

$$\begin{aligned}\widehat{\text{Var}}\left(\widehat{S}(t)\right) &= -\frac{\widehat{S}(t)^2}{n} \int_0^t \frac{d\widehat{H}_1(u)}{\widehat{H}(u)\widehat{H}(u-)} \\ &= \frac{\widehat{S}(t)^2}{n} \frac{1}{n} \sum_{i=1}^n \frac{D_i \mathbb{1}_{\{Y_i \leq t\}}}{\frac{1}{n^2} \left(\sum_{j=1}^n \mathbb{1}_{\{Y_j > Y_i\}} \right) \left(\sum_{j=1}^n \mathbb{1}_{\{Y_j \geq Y_i\}} \right)} \\ &= \widehat{S}(t)^2 \sum_{i=1}^n \frac{D_i \mathbb{1}_{\{Y_i \leq t\}}}{\left(\sum_{j=1}^n \mathbb{1}_{\{Y_j > Y_i\}} \right) \left(\sum_{j=1}^n \mathbb{1}_{\{Y_j \geq Y_i\}} \right)}\end{aligned}$$

2. En notant $Y_{(i)}$ la statistique d'ordre i des Y et $D_{[i]}$ la valeur de D associée, ainsi que $r_{[i]}$ le nombre d'individus à risque juste avant la date $T_{(i)}$ et $d_{[i]}$ le nombre de décès à cette date, on obtient

$$\widehat{\text{Var}}\left(\widehat{S}(t)\right) = \widehat{S}(t)^2 \sum_{i: T_{(i)} \leq t} \frac{d_{[i]}}{(r_{[i]} - d_{[i]}) r_{[i]}} = \widehat{S}(t)^2 \gamma(t)^2.$$

On reconnaît l'estimateur de Greenwood (qui est consistant).

3. On obtient un intervalle de confiance asymptotique d'ordre α pour l'estimateur de Kaplan-Meier

$$\left[\widehat{S}(t) \left(1 - \phi^{-1} \left(1 - \frac{\alpha}{2} \right) \gamma(t) \right), \widehat{S}(t) \left(1 + \phi^{-1} \left(1 - \frac{\alpha}{2} \right) \gamma(t) \right) \right],$$

avec ϕ la fonction de répartition d'un loi normale centrée réduite.

4. En remarquant que

$$\widehat{p}_t = \prod_{i: t < Y_i \leq t+1} \left(1 - \frac{D_i}{\sum_{j=1}^n \mathbb{1}_{\{Y_i \leq Y_j\}}} \right) = \prod_{i: t < Y_{(i)} \leq t+1} \left(1 - \frac{d_{[i]}}{r_{[i]}} \right),$$

on en déduit facilement que

$$\widehat{\text{Var}}(\widehat{p}_t) = \widehat{p}_t^2 \sum_{i: t < Y_{(i)} \leq t+1} \frac{d_{[i]}}{(r_{[i]} - d_{[i]}) r_{[i]}}.$$

Exercice 3 Modèle à risques concurrents.

Soit T la durée de vie d'un individu que l'on suppose soumise à K causes de sortie (ex : causes de décès). On définit la *fonction de hasard spécifique au risque i*

$$h^{(i)}(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \mathbb{P}(t < T \leq t + \Delta t, V = i | T > t),$$

où V est la variable aléatoire qui désigne la cause de sortie.

1. Donner l'expression de :

- la fonction de hasard h (associée à T) en fonction des $h^{(i)}$;
- la fonction de survie S (associée à T) en fonction des $h^{(i)}$;

- la fonction d'incidence cumulée $F^{(i)}(t) = \mathbb{P}(T \leq t, V = i)$;
 - le taux de décès entre les dates t et $t + 1$ pour la cause $V = i$.
2. On introduit les durées de vie latentes par cause $T_1, T_2, \dots, T_K$. Exprimer $S(t)$ en fonction des fonctions de survie de chacune des lois latentes en faisant l'hypothèse qu'elles sont indépendantes. Que peut-on en déduire s'agissant de l'expression des taux de hasard de chacune de ces lois latentes ?
3. En notant $S_{(T_1, \dots, T_K)}(t_1, \dots, t_K)$ la fonction de survie jointe de $T_1, T_2, \dots, T_K$ et sans supposer que ces lois sont nécessairement indépendantes, montrer que

$$h^{(i)}(t) = -\frac{1}{S(t)} \frac{\partial S_{(T_1, \dots, T_K)}(t_1, \dots, t_K)}{\partial t_i} \Big|_{t_1, \dots, t_K=t}.$$

4. Pour $K = 2$, nous faisons à présent l'hypothèse que les durées de vie latentes sont corrélées et reliées par la fonction de survie jointe

$$S_{T_1, T_2}(t_1, t_2) = (1 + \theta(\lambda_1 t_1 + \lambda_2 t_2))^{-\frac{1}{\theta}},$$

avec $\lambda_1, \lambda_2, \theta \geq 0$. Fournir l'expression de la fonction de hasard $h^{(i)}(t)$ pour $i = 1, 2$, puis celle de $F^{(i)}(t)$.

5. On se place à présent dans un cadre d'estimation non-paramétrique avec censure à droite indépendante (C indépendant de T et de V), similaire à celui de l'exercice 1 (même notation et par simplicité on utilise désormais la notation T_i pour la loi du durée de l'individu i). On souhaite proposer un estimateur non-paramétrique pour $F^{(i)}(t)$. En notant $H_1^{(i)}(t) = \mathbb{P}(Y > t, D = 1, V = i)$ et $H(t) = \mathbb{P}(Y > t)$, montrer que

$$\frac{dF^{(i)}(t)}{S(t-)} = -\frac{dH_1^{(i)}(t)}{H(t-)}.$$

Conclure en suivant des raisonnements similaires à ceux de l'exercice 1.

Réponse de l'exercice 3.

1. La fonction de hasard s'obtient par

$$h(t) = \sum_{i=1}^K h^{(i)}(t).$$

La fonction de survie s'écrit

$$S(t) = \exp\left(-\int_0^t h(u) du\right) = \exp\left(-\sum_{i=1}^K \int_0^t h^{(i)}(u) du\right).$$

La fonction d'incidence cumulée s'écrit

$$F^{(i)}(t) = \int_0^t h^{(i)}(u) S(u) du = \int_0^t h^{(i)}(u) \exp\left(-\sum_{i=1}^K \int_0^u h^{(i)}(s) ds\right) du.$$

Le taux de décès est

$$\begin{aligned} {}_1q_t^{(i)} &= \mathbb{P}(t < T \leq t+1, V = i | T > t) \\ &= \frac{\int_t^{t+1} h^{(i)}(u) S(u) du}{S(t)} = \frac{F^{(i)}(t+1) - F^{(i)}(t)}{S(t)}. \end{aligned}$$

2. On a $T = \min_{i=1,\dots,K} (T_i)$. Par conséquent sous l'hypothèse d'indépendance, la fonction de survie s'écrit en fonction des fonctions de survie des lois marginales $S_i(t)$

$$S(t) = \prod_{i=1}^K \mathbb{P}(T_i > t) = \prod_{i=1}^K S_i(t).$$

On montre également sous cette hypothèse que

$$\begin{aligned} F^{(i)}(t) &= \mathbb{P}\left(T_i \leq t, T_i \leq \min_{j \neq i} (T_j)\right) \\ &= \mathbb{E}\left[\mathbb{1}_{\{T_i \leq t\}} \mathbb{E}\left[\mathbb{1}_{\{T_i \leq \min_{j \neq i} (T_j)\}} | T_i\right]\right] \\ &= \int_0^t \prod_{j \neq i} S_j(u) f_i(u) du, \end{aligned}$$

avec f_i la densité de T_i .

Dans ces conditions, la fonction de hasard spécifique au risque i s'écrit

$$\begin{aligned} h^{(i)}(t) &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \mathbb{P}(t < T \leq t + \Delta t, V = i | T > t) \\ &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \Delta t q_t^{(i)} \\ &= \frac{1}{S(t)} \frac{dF^{(i)}(t)}{dt} = \frac{1}{S(t)} \prod_{j \neq i} S_j(t) f_i(t) \\ &= \frac{f_i(t)}{S_i(t)} = h_i(t) \end{aligned}$$

avec h_i la fonction de hasard de T_i .

3. On écrit que (avec des notations évidentes)

$$\begin{aligned} &\left. \frac{\partial S_{(T_1, \dots, T_K)}(t_1, \dots, t_K)}{\partial t_i} \right|_{t_1, \dots, t_K = t} \\ &= -f_i(t) \int_t^\infty \cdots \int_t^\infty dF_{(T_1, \dots, T_{i-1}, T_{i+1}, \dots, T_K) | T_i}(t_1, \dots, t_{i-1}, t_{i+1}, \dots, t_K | t) \\ &= -\frac{dF^{(i)}(t)}{dt}. \end{aligned}$$

D'où le résultat.

4. En reprenant le résultat précédent et comme $S(t) = S_{(T_1, T_2)}(t, t)$, on a pour $i = 1, 2$

$$h^{(i)}(t) = -\frac{1}{S(t)} \left. \frac{\partial S_{(T_1, T_2)}(t_1, t_2)}{\partial t_i} \right|_{t_1, t_2 = t} = \frac{\lambda_i}{(1 + \theta t (\lambda_1 + \lambda_2))},$$

et

$$\begin{aligned} F^{(i)}(t) &= \int_0^t \frac{\lambda_i}{(1 + \theta u (\lambda_1 + \lambda_2))} (1 + \theta u (\lambda_1 + \lambda_2))^{-\frac{1}{\theta}} du \\ &= \frac{\lambda_i}{(\lambda_1 + \lambda_2)} \left(1 - (1 + \theta t (\lambda_1 + \lambda_2))^{-\frac{1}{\theta}}\right). \end{aligned}$$

5. On écrit

$$\begin{aligned}
H_1^{(i)}(t) &= \mathbb{E} [\mathbb{1}_{\{T>t, V=i\}} \mathbb{1}_{\{C>T\}}] \\
&= \mathbb{E} [\mathbb{1}_{\{T>t, V=i\}} \mathbb{E} [\mathbb{1}_{\{C>T\}} | T, V]] \\
&= \mathbb{E} [\mathbb{1}_{\{T>t, V=i\}} G(T-)] \\
&= \int_t^\infty G(u-) dF^{(i)}(u).
\end{aligned}$$

On a donc directement par les mêmes calculs que l'exercice 1

$$\frac{dF^{(i)}(t)}{S(t-)} = -\frac{dH_1^{(i)}(t)}{H(t-)}.$$

Un premier résultat (avec les mêmes notations) est que

$$\widehat{F}^{(i)}(t) = \sum_{i=1}^n W_{i,n} \mathbb{1}_{\{V=i\}} \mathbb{1}_{\{Y_i \leq t\}}.$$

On a également

$$\widehat{H}_1^{(i)}(t) = \frac{1}{n} \sum_{i=1}^n D_i \mathbb{1}_{\{Y_i > t, V=i\}}.$$

Un second résultat s'obtient en considérant l'estimateur de Kaplan-Meier pour S tel que

$$\begin{aligned}
\widehat{F}^{(i)}(t) &= - \int_0^t \widehat{S}(t-) \frac{d\widehat{H}_1^{(i)}(t)}{\widehat{H}(t-)} \\
&= \frac{1}{n} \sum_{j=1}^n \frac{D_j \widehat{S}(Y_j-) \mathbb{1}_{\{Y_j \leq t, V_j=i\}}}{\widehat{H}(Y_j-)} \\
&= \sum_{i=1}^n \frac{D_j \widehat{S}(Y_j-) \mathbb{1}_{\{Y_j \leq t, V_j=i\}}}{\sum_{k=1}^n \mathbb{1}_{\{Y_k \geq Y_j\}}} \\
&= \sum_{Y_j \leq t} \frac{D_j \widehat{S}(Y_j-) \mathbb{1}_{\{V_j=i\}}}{\sum_{k=1}^n \mathbb{1}_{\{Y_k \geq Y_j\}}}.
\end{aligned}$$

En ordonnant sur les valeurs de Y_i , on obtient

$$\widehat{F}^{(i)}(t) = \sum_{Y_{(j)} \leq t} \frac{D_{[j]} \mathbb{1}_{\{V_{[j]}=i\}}}{n-j+1} \widehat{S}(Y_{[j]}-).$$