
Travaux dirigés : modèles de durée
Séance n°1 - Corrigé

Exercice 1.

Soit T une variable aléatoire positive représentant la durée de vie d'un individu. Soit S sa fonction de survie.

- Supposons que T soit une variable aléatoire continue. Donner l'expression de sa fonction de survie, de sa densité f , puis trouver la relation qui les lient à la fonction de hasard définie par

$$h(t) = \lim_{\delta t \rightarrow 0} \frac{\mathbb{P}(t < T \leq t + \delta t \mid T > t)}{\delta t}.$$

- Montrer que

$$S(t) = \exp(-H(t)),$$

avec $H(t) = \int_0^t h(u) du$, appelée fonction de hasard cumulée.

- Supposons à présent que T soit discrète et prenne ses valeurs aux dates $t_1 < \dots < t_k$. En notant, $p(t_i) = \mathbb{P}(T = t_i)$, pour $i = 1, \dots, k$, fournir une expression additive et multiplicative de $S(t)$.
- En temps discret, on définit usuellement la fonction de hasard par $h(t_i) = \frac{p(t_i)}{S(t_{i-1})}$ avec $S(t_0) = 1$. En déduire, une expression de S en fonction de h .
- Comparer cette expression au cas continu. Que faire pour que la relation de la question 2. s'applique au cas discret.

Réponse de l'exercice 1.

- On a :

- $S(t) = \mathbb{P}(T > t);$
- $f(t) = -\frac{d}{dt}S(t);$
- $h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \ln(S(t)).$

- Il s'agit de la solution d'une équation différentielle d'ordre 1 avec $S(0) = 1$.

- Par définition, on a

$$S(t) = \sum_{t_i > t} p(t_i).$$

On peut aussi noter

$$S(t) = \prod_{t_i \leq t} \frac{S(t_i)}{S(t_{i-1})}.$$

4. En remarquant que $p(t_i) = S(t_{i-1}) - S(t_i)$, on a immédiatement que $h(t_i) = 1 - \frac{S(t_i)}{S(t_{i-1})}$.
D'où,

$$S(t) = \prod_{t_i \leq t} (1 - h(t_i)) = \prod_{t_i \leq t} (1 - \Delta H(t_i)),$$

avec $H(t) = \sum_{t_i \leq t} h(t_i)$. Il est utile de conserver cette expression en tête lorsque l'on examinera l'estimateur de Kaplan-Meier de la fonction de survie (cours sur les méthodes d'estimation non-paramétriques).

5. Certains auteurs préfèrent définir H en temps discret telle que

$$H(t) = - \sum_{t_i \leq t} \ln(1 - h(t_i)),$$

pour bénéficier de la relation de la question 2.

Exercice 2.

Soit T une variable aléatoire positive continue représentant la durée de vie d'un individu. Soit S sa fonction de survie, f sa densité et h sa fonction de hasard.

1. Donner les expressions, en fonction de $S(t)$, de l'espérance de vie de T et de l'espérance de vie résiduelle de T à la date x

$$\mathbb{E}[T - x \mid T > x].$$

2. Donner l'expression de la variance de T .

3. En considérant un contrat d'assurance versant un montant de 1 EUR à la date du décès de l'assuré et un taux d'actualisation constant δ (en temps continu), donner l'expression de la valeur de l'engagement de l'assuré d'âge x .

Réponse de l'exercice 2.

1. On a $\mathbb{E}[T] = \mathbb{E}\left[\int_0^T dt\right] = \int_0^\infty \mathbb{E}[\mathbb{1}_{\{T>t\}}] dt = \int_0^\infty S(t) dt$.

On applique le même raisonnement avec $T_x = (T - x) \mathbb{1}_{\{T>x\}}$ de fonction de survie $S_x(t) = S(t+x)/S(x)$, et donc

$$\mathbb{E}[T - x \mid T > x] = \int_0^\infty S_x(t) dt = \int_x^\infty \frac{S(t)}{S(x)} dt.$$

2. Par définition, $\mathbb{V}(T) = \mathbb{E}(T^2) - (\mathbb{E}(T))^2$. Par changement de variable et le même raisonnement que précédemment, on a

$$\mathbb{V}(T) = 2 \int_0^\infty t S(t) dt - \left(\int_0^\infty S(t) dt \right)^2.$$

On a un résultat analogue pour la durée de survie résiduelle.

3. L'engagement de l'assureur est représenté par la variable $\Lambda_x = e^{\delta T_x}$. D'où l'on tire (avec des notations évidentes) :

— la VAP de l'engagement $\bar{A}_x = \mathbb{E}[\Lambda_x] = \int_x^\infty e^{\delta t} f_x(t) dt = \int_x^\infty e^{\delta t} h_x(t) S_x(t) dt$;

— la variance de l'engagement $\mathbb{V}[\Lambda_x] = \int_x^\infty e^{2\delta t} h_x(t) S_x(t) dt - (\bar{A}_x)^2 =^2 \bar{A}_x - (\bar{A}_x)^2$, où ${}^2\bar{A}_x$ est la VAP de l'engagement calculée avec un taux de 2δ .

Cette relation, souvent appelée *règle des moments*, se généralise à des moments d'ordre supérieur et à d'autres garanties.

Exercice 3 Modèle à hasard constant par morceaux.

Pour une décomposition en K segments fixés de l'ensemble de valeurs prises par la durée de vie T , la fonction de hasard du modèle est supposée constante par morceaux (appelé aussi modèle de Poisson) telle que

$$h(t) = \theta_k \text{ pour } t \in J_k = [\tau_{k-1}; \tau_k[, \quad k = 1, \dots, K, \text{ avec } \tau_0 = 0 \text{ et } \tau_K = \infty,$$

où $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ est un vecteur de paramètres positifs.

1. Si $K = 1$, que dire de la loi suivie par la variable T .
2. Si $K > 1$, donner l'expression de la fonction de survie de T .
3. Calculer l'espérance de vie résiduelle à un âge $x \in J_l$ avec $l \in \{1, \dots, K\}$.
4. Soit un échantillon de n individus i.i.d. de durée de vie T_1, \dots, T_n . Écrire la log-vraisemblance du modèle.
5. Calculer le score du modèle et en déduire un estimateur pour chaque θ_k en faisant apparaître pour chaque segment k une statistique comptant le nombre de décès N_k et une autre mesurant l'exposition au risque R_k .
6. Calculer la matrice d'information de Fisher. En déduire une expression de la variance asymptotique des $\hat{\theta}_k$ et proposer un intervalle de confiance asymptotique de niveau α pour ces estimateurs.
7. Expliciter un test pour l'hypothèse $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Réponse de l'exercice 3.

1. Pour $K = 1$, la fonction de hasard est constante sur $[0; \infty[$, i.e. $h(t) = \theta$ pour $t \geq 0$. On a donc $S(t) = \exp(-\theta t)$, ce qui correspond à une loi exponentielle de paramètre $\theta \geq 0$.
2. On rappelle que $S(t) = \exp\left(-\int_0^t h(u) du\right)$. Comme

$$\int_0^t h(u) du = \sum_{k=1}^K (\tau_k \wedge t - \tau_{k-1})^+ \theta_k = \sum_{k=1}^K r_k(t) \theta_k,$$

avec $r_k(t)$ le temps passé dans le segment k , on en déduit que $S(t) = \exp\left(-\sum_{k=1}^K r_k(t) \theta_k\right)$. Autrement dit, si $t \in J_l$, on a

$$S(t) = \exp\left(-\sum_{k=1}^{l-1} \theta_k (\tau_k - \tau_{k-1}) - \theta_l (t - \tau_{l-1})\right)$$

3. On suppose que $x \in J_l$. On note que

$$S_x(t) = \frac{S(x+t)}{S(x)} = \exp \left(-(\tau_l \wedge (x+t) - x) \theta_l - \sum_{k=l+1}^K r_k(t+x) \theta_k \right).$$

On obtient

$$\begin{aligned} \mathbb{E}_x[T_x] &= \mathbb{E}[T - x \mid T > x] \\ &= \frac{1}{S(x)} \left(\int_x^{\tau_l} S(t) dt + \sum_{k=l+1}^K \int_{\tau_{k-1}}^{\tau_k} S(t) dt \right) \\ &= \int_x^{\tau_l} e^{-\theta_l(t-x)} dt + \int_{\tau_l}^{\tau_{l+1}} e^{-\theta_l(\tau_l-x)-\theta_{l+1}(t-\tau_l)} dt \\ &\quad + \sum_{k=l+2}^K \int_{\tau_{k-1}}^{\tau_k} e^{-\theta_l(\tau_l-x)-\sum_{j=l+1}^{k-1} \theta_j(\tau_j-\tau_{j-1})-\theta_k(t-\tau_{k-1})} dt \\ &= \frac{1 - e^{-\theta_l(\tau_l-x)}}{\theta_l} + \frac{e^{-\theta_l(\tau_l-x)} (1 - e^{-\theta_{l+1}(\tau_{l+1}-\tau_l)})}{\theta_{l+1}} \\ &\quad + e^{-\theta_l(\tau_l-x)} \sum_{k=l+1}^K \theta_k^{-1} e^{-\sum_{j=l+1}^{k-1} \theta_j(\tau_j-\tau_{j-1})} (1 - e^{-\theta_k(\tau_k-\tau_{k-1})}). \end{aligned}$$

4. On observe $(t_i)_{i=1,\dots,n}$. La log-vraisemblance s'écrit

$$\begin{aligned} \ln(\mathcal{L}(\boldsymbol{\theta})) &= \sum_{i=1}^n \ln(f(t_i)) \\ &= \sum_{i=1}^n \ln(h(t_i)) - \int_0^{t_i} h(u) du. \end{aligned}$$

En notant $D_{i,k} = \mathbb{1}_{\{T_i \in J_k\}}$, on a

$$\ln(h(t_i)) = \sum_{k=1}^K d_{i,k} \ln(\theta_k),$$

et

$$\int_0^{t_i} h(u) du = \sum_{k=1}^K r_k(t_i) \theta_k,$$

Ainsi, on en déduit

$$\ln(\mathcal{L}(\boldsymbol{\theta})) = \sum_{k=1}^K N_k \ln(\theta_k) - \sum_{k=1}^K R_k \theta_k,$$

avec $N_k = \sum_i^n d_{i,k}$ le nombre de décès observés dans le segment k et $R_k = \sum_i^n r_k(t_i)$ le temps d'exposition total (ou exposition au risque) de la population dans ce même segment.

Remarque : l'équation de la log-vraisemblance serait équivalente (à une constante près) à celle obtenue avec une modélisation des nombres de décès N_k selon une loi de Poisson telle que

$$N_k \sim \mathcal{P}(R_k \theta_k).$$

On parle donc usuellement de modèle de Poisson pour le nombre de décès.

5. En dérivant par rapport à θ_k , la k -ième composante du vecteur de score s'écrit

$$U_k(\theta) = \frac{\partial}{\partial \theta_k} \ln(\mathcal{L}(\theta)) = \frac{N_k}{\theta_k} - R_k,$$

et il vient

$$\hat{\theta}_k = \frac{N_k}{R_k}.$$

Cet estimateur correspond au ratio d'un nombre de décès ramené à une exposition au risque.

6. En dérivant une seconde fois, le terme situé en position (k, l) de la matrice d'information de Fisher vaut

$$\mathcal{I}_{kl}(\theta) = \delta_{kl} \frac{N_k}{\theta_k \theta_l},$$

avec δ_{kl} le symbole de Kronecker. On en déduit directement la variance asymptotique de $\hat{\theta}_k$, puis une expression d'un intervalle de confiance asymptotique

$$\left[\frac{N_k}{R_k} - \phi^{-1} \left(1 - \frac{\alpha}{2} \right) \frac{\sqrt{(N_k)}}{R_k}, \frac{N_k}{R_k} + \phi^{-1} \left(1 - \frac{\alpha}{2} \right) \frac{\sqrt{(N_k)}}{R_k} \right],$$

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

7. Trois tests classiques peuvent être proposés pour tester l'hypothèse $\theta = \theta_0$:

- statistique de Wald : $(\hat{\theta} - \theta_0)^\top \mathcal{I}(\hat{\theta}) (\hat{\theta} - \theta_0)$;
- statistique de score : $\mathbf{U}(\theta_0)^\top \mathcal{I}(\theta_0)^{-1} \mathbf{U}(\theta_0)$;
- ratio des vraisemblance : $2 \left(\ln \mathcal{L}(\hat{\theta}) - \ln \mathcal{L}(\theta_0) \right)$.

Explicitons par exemple la statistique de Wald

$$\begin{aligned} (\hat{\theta} - \theta_0)^\top \mathcal{I}(\hat{\theta}) (\hat{\theta} - \theta_0) &= \sum_{k=1}^K (\hat{\theta}_k - \theta_{0,k})^2 \mathcal{I}_{kk}(\hat{\theta}) \\ &= \sum_{k=1}^K \left(\frac{N_k}{R_k} - \theta_{0,k} \right)^2 \frac{R_k^2}{N_k}. \end{aligned}$$

Exercice 4 Modèles AFT et avec *odds* proportionnels.

On souhaite étudier deux classes de modèles classiques de régression : les modèles *accelerated failure times* (AFT), traduisant une multiplication de la durée de vie par rapport à une durée de référence, et les modèles avec *odds* proportionnels (PO), traduisant une multiplication de l'*odds* des fonctions de répartition (ou de survie). Dans la suite, on introduit \mathbf{X} un vecteur de covariables.

1. On considère un modèle AFT tel que la loi de durée prend la forme

$$\ln T = -\mathbf{X}^\top \boldsymbol{\theta} + W,$$

avec $\boldsymbol{\theta}$ un vecteur de paramètres et une distribution correspondant à un terme d'erreur. Écrire la loi de survie de T , $S(t|\mathbf{X}; \boldsymbol{\theta})$, et sa fonction de hasard, $h(t|\mathbf{X}; \boldsymbol{\theta})$, en fonction de celles de la loi de référence $T_0 = \exp(W)$.

2. Montrer que si T_0 suit une loi de Weibull, de fonction de hasard

$$h_0(t) = \lambda \alpha t^{\alpha-1}, \lambda > 0, \alpha > 0,$$

alors la loi du modèle AFT est encore une loi de Weibull. Donner sa fonction de survie.

3. Les modèles PO sont définis par la relation générale

$$\frac{1 - S(t)}{S(t)} = \frac{1 - S_0^*(t)}{S_0^*(t)} \exp(\mathbf{X}^\top \boldsymbol{\beta}),$$

avec $S_0^*(t)$ la fonction de survie d'une loi quelconque et $\boldsymbol{\beta}$ un vecteur de paramètres. Montrer que si T_0 suit une loi de Weibull, alors le modèle AFT ne satisfait pas la relation d'un modèle PO.

4. Caractériser la loi de S_0 pour que le modèle AFT vérifie l'hypothèse PO. En prenant différentes valeurs de \mathbf{X} , on cherchera à maintenir le ratio $\frac{\beta_j}{\theta_j}$ constant.

Réponse de l'exercice 4.

1. En passant à l'exponentielle, on a

$$T = \exp(W) \exp(-\mathbf{X}^\top \boldsymbol{\theta}) = T_0 \exp(-\mathbf{X}^\top \boldsymbol{\theta}).$$

D'où l'on tire immédiatement

$$S(t|\mathbf{X}; \boldsymbol{\theta}) = \mathbb{P}(T_0 \exp(-\mathbf{X}^\top \boldsymbol{\theta}) > t) = S_0(t \exp(\mathbf{X}^\top \boldsymbol{\theta})),$$

et

$$h(t|\mathbf{X}; \boldsymbol{\theta}) = h_0(t \exp(\mathbf{X}^\top \boldsymbol{\theta})) \exp(\mathbf{X}^\top \boldsymbol{\theta}).$$

2. Si T_0 suit une loi de Weibull, alors par construction, on a, en notant $\gamma = \exp(\mathbf{X}^\top \boldsymbol{\theta})$

$$h(t|\mathbf{X}; \boldsymbol{\theta}) = h_0(t\gamma) \gamma = \lambda \alpha (t\gamma)^{\alpha-1} \gamma = \lambda^* \alpha t^{\alpha-1},$$

avec $\lambda^* = \lambda \gamma^\alpha$. La fonction de survie est donc

$$S(t|\mathbf{X}; \boldsymbol{\theta}) = \exp(-\lambda^* t^\alpha).$$

3. On cherche à présent un cas particulier permettant de montrer que l'hypothèse PO n'est pas vérifiée. En effet, si elle était vérifiée, on aurait

$$\frac{1 - S(t|\mathbf{X}; \boldsymbol{\theta})}{S(t|\mathbf{X}; \boldsymbol{\theta})} = \frac{1 - S_0^*(t)}{S_0^*(t)} \exp(\mathbf{X}^\top \boldsymbol{\beta}).$$

Cette égalité serait encore vraie pour $\mathbf{X} = 0$ et comme la fonction $x \mapsto \frac{1-x}{x}$ est strictement décroissante, on aurait

$$S_0^*(t) = S_0(t) = \exp(-\lambda t^\alpha).$$

En prenant $\mathbf{X} = (1, 0, 0, \dots)$, l'égalité deviendrait

$$\frac{1 - \exp(-\lambda \exp(\alpha \theta_1) t^\alpha)}{\exp(-\lambda \exp(\alpha \theta_1) t^\alpha)} = \frac{1 - \exp(-\lambda t^\alpha)}{\exp(-\lambda t^\alpha)} \exp(\beta_1).$$

Celle-ci n'est pas vérifiée pour tout $t \geq 0$, d'où le résultat.

4. On chercher à présent à caractériser la loi S_0 comme intersection entre un modèle AFT et un modèle PO. Comme dans la question précédente, on montre que

$$S_0^*(t) = S_0(t).$$

En prenant $\mathbf{X} = \left(-\frac{1}{\theta_1} \ln(t), 0, 0, \dots\right)$, l'égalité devient

$$\frac{1 - S_0(1)}{S_0(1)} = \frac{1 - S_0^*(t)}{S_0^*(t)} \exp\left(-\frac{\beta_1}{\theta_1} \ln(t)\right),$$

et donc pour tout $t \geq 0$

$$\frac{1 - S_0(t)}{S_0(t)} = \frac{1 - S_0(1)}{S_0(1)} t^{\frac{\beta_1}{\theta_1}}.$$

En appliquant le même raisonnement avec $X_j = -\frac{1}{\theta_j} \ln(t)$, on en déduit que le ratio des coefficients doit être constant

$$\frac{\beta_j}{\theta_j} = p.$$

D'où

$$\frac{1 - S_0(t)}{S_0(t)} = \frac{1 - S_0(1)}{S_0(1)} t^p = (ct)^p,$$

et donc

$$S_0(t) = \frac{1}{1 + (ct)^p}.$$

On reconnaît la fonction de survie d'une loi log-logistique.

Exercice 5 Fragilité Gamma.

Soit T la variable aléatoire positive et continue représentant la durée de vie d'un individu de fonction de survie S . Pour une population donnée, on cherche à mesurer l'effet d'une source d'hétérogénéité latente en date $t = 0$, représentée par une variable Z de densité $\pi(z)$. Les individus pour lesquels $Z = z$ sont supposés suivre la même loi de durée, de fonction de survie $S(t|z)$, de densité $f(t|z)$ et fonction de hasard

$$h(t|z) = zh(t).$$

La variable Z est usuellement appelée *fragilité*.

- Après avoir écrit $S(t)$ en fonction de $S(t|z)$, donner l'expression de $\pi_t(z)$, la densité de la fragilité pour la population des survivants en date $t \geq 0$, i.e. sachant $T \geq t$. Dans la suite, on notera $Z_t = Z \mid T \geq t$ la variable de densité $\pi_t(z)$.

2. On suppose que la fragilité suit initialement une loi Gamma (λ, k) de densité*

$$\pi(z) = \frac{\lambda^k z^{k-1} \exp(-\lambda z)}{\Gamma(k)}, k > 0, \lambda > 0.$$

Donner l'expression de $S(t)$.

3. Fournir l'expression de la fonction de hasard moyenne (non conditionnelle) pour la population survivante en $t \geq 0$

$$\bar{h}(t) = \int_0^\infty h(t|z) \pi_t(z) dz.$$

4. Calculer l'espérance de Z_t , puis dériver la. Commenter le résultat.
 5. Proposer un paramétrage pour k et λ permettant d'interpréter facilement la fonction de hasard des survivants $\bar{h}(t)$, ainsi que l'espérance et la variance de Z_t .

Réponse de l'exercice 5.

1. Avec le temps, les proportions d'individus prenant une même valeur pour Z se modifie. On note tout d'abord que $S(t) = \int S(t|z) \pi(z) dz$. Ainsi, on a pour la population des survivants en date t

$$\pi_t(z) = \frac{S(t|z) \pi(z)}{\int S(t|z) \pi(z) dz} = \frac{S(t|z)}{S(t)} \pi(z).$$

La densité de Z_t est multipliée par la proportion de survivants dans chaque sous-groupe, caractérisé initialement par la même valeur de Z .

2. Comme $h(t|z) = zh(t)$, on a

$$\begin{aligned} S(t|z) \pi(z) &= \exp(-zH(t)) \frac{\lambda^k z^{k-1} \exp(-\lambda z)}{\Gamma(k)} \\ &= \frac{\lambda^k}{(\lambda(t))^k} \frac{(\lambda(t))^k z^{k-1} \exp(-\lambda(t)z)}{\Gamma(k)}, \end{aligned}$$

avec $\lambda(t) = \lambda + H(t)$. En isolant le terme $\frac{\lambda^k}{(\lambda(t))^k}$, on reconnaît une loi Gamma de paramètre $(k, \lambda(t))$. D'où,

$$S(t) = \frac{\lambda^k}{(\lambda(t))^k}.$$

3. Le terme $\frac{\lambda^k}{(\lambda(t))^k}$ se simplifie au numérateur et au dénominateur de $\pi_t(z)$, qui correspond alors à la densité d'une loi Gamma de paramètre $(k, \lambda(t))$. La fonction de hasard moyenne s'écrit donc

$$\bar{h}(t) = \int_0^\infty h(t|z) \pi_t(z) dz = h(t) \int_0^\infty z \pi_t(z) dz = h(t) \frac{k}{\lambda(t)}.$$

On note que h croît plus vite que \bar{h} , puisque H est croissante.

*. Rappel : $\mathbb{E}[Z] = k/\lambda$ et $\mathbb{V}[Z] = k/\lambda^2$.

4. L'espérance de Z_t s'écrit

$$\int_0^\infty z\pi_t(z)dz = \frac{k}{\lambda(t)} = \bar{z}(t).$$

On a alors

$$\frac{d}{dt}\bar{z}(t) = -h(t)\frac{k}{\lambda(t)^2} = -h(t)\sigma(t)^2,$$

avec $\sigma(t)^2$ la variance de Z_t . On observe que la moyenne de la fragilité décroît avec le temps, puisque les décès conduisent à toucher plus tôt les individus pour lesquels la fragilité est la plus grande.

5. En supposant que $\lambda = k = \frac{1}{\sigma^2}$ (σ^2 est la variance de Z), on obtient

$$\bar{h}(t) = h(t)S(t)^{\sigma^2} = \frac{h(t)}{1 + \sigma^2 H(t)},$$

$$\mathbb{E}(Z_t) = \frac{1}{1 + \sigma^2 H(t)},$$

et

$$\mathbb{V}(Z_t) = \sigma(t)^2 = \frac{\sigma^2}{(1 + \sigma^2 H(t))^2}.$$

L'espérance de la fragilité des survivants décroît avec le temps. C'est aussi le cas pour la variance, ce qui signifie d'une certaine manière que la population devient plus homogène avec le temps. Cependant, le coefficient de variation $\sqrt{\mathbb{V}(Z_t)}/\mathbb{E}(Z_t)$ est constant ce qui signifie que d'un point de vue relatif, l'hétérogénéité de la population demeure la même.