EXAMEN FINAL

Modèle de Durée– 2022-2023
Pierre-O Goffard

Instructions: On éteint et on range son téléphone.

- La calculatrice et les appareils éléctroniques ne sont pas autorisés.
- Vous devez justifier vos réponses de manière claire et concise.
- Vous devez écrire de la manière la plus lisible possible. Souligner ou encadrer votre réponse finale.
- Document autorisé: Une feuille manuscrite recto-verso

Question:	1	2	Total
Points:	5	15	20
Score:			

1. Soient deux variables aléatoires $E \sim \text{Exp}(\lambda)$ et $W \sim \text{Weibull}(\alpha, \beta)$, mutuellement indépendantes, de densité respective

$$f_E(x) = \lambda e^{-\lambda x}, \text{ et } f_W(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-(x/\beta)^\alpha},$$

pour $x > 0$. On définit également $T = \min(E, W)$, il s'agit d'un modèle de risques concurrents.

(a) (2 points) Rappeler la définition de la fonction de hasard, et donner l'expression des fonctions de hasard de E et W.

Solution: La fonction de hasard est défini par

$$h(x) = \frac{f(x)}{S(x)},$$

où f et S sont la densité et la fonction de survie respectivement. Pour les deux modèle considéré on a

$$h_E(x) = \lambda, \text{ et } h_W(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}.$$

(b) (1 point) Calculer $S(t) = \mathbb{P}(T > t)$, la fonction de survie de T.

Solution:

$$S(t) = \mathbb{P}(T > t) = \mathbb{P}(\min(E, W) > t) = S_E(t)S_W(t) = \exp(-\lambda x - (x/\beta)^\alpha)$$

(c) (1 point) Calculer $h(t)$ la fonction de hasard de T.

Solution: On a

$$f(t) = S'(t) = f_E(t)S_W(t) + S_E(t)f_W(t).$$

On en déduit que

$$h(t) = h_E(t) + h_W(t) = \lambda + \frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1}$$

(d) (1 point) Nous disposons d'un jeu de n observations iid, notées $t_1, \ldots, t_n$, de T censurées à droite. Donner l'expression de la vraisemblance du modèle pour ces données censurées. Vous devez rappeler les notations utilisées habituellement.

Solution: Soient $t_1, \ldots, t_n$, les données disponibles sont données par

$$\mathcal{D} = (x_i, \delta_i) = (t_i \wedge c_i, \mathbb{I}_{t_i \leqslant c_i})$$

où $c_1, \ldots, c_n$ sont les réalisations de la variable de censure. La vraisemblance s'écrit

$$\mathcal{L}(\mathcal{D}, \theta) = \prod_{i=1}^{n} h(x_k)^{\delta_i} S(x_k),$$

où $\theta = (\lambda, \alpha, \beta)$.

2. Le tableau 1 est un extrait d'un jeu de données comprenant les expositions initiales et nombre de décès dans la population Néerlandaise pour l'année 2000.

Age	E_x^0	D_x
0	204,776	1,059
1	202,369	94
2	196,887	59
⋮	⋮	
108	3	2
109	1	0
110+	3	0

Table 1: Extrait des données de mortalités au Pays-Bas.

(a) (2 points) Donner l'expression des probabilités de décès à l'âge x notées $\widehat{q}_x$ en fonction de E_x^0. Rappeler les hypothèses du modèle sous jacent et à quoi correspond E_x^0.

Solution: E_x^0 est l'exposition initiale. On utilise le modèle binomial suivant lequel le nombre de décès à l'âge x suit une loi binomiale

$$D_x \sim \text{Bin}(E_x^0, q_x).$$

Dans le cadre de ce modèle, on estime la probabilité de décès par

$$\widehat{q_x} = \frac{D_x}{E_x^0}.$$

(b) (2 points) Les probabilités de décès "brutes" sont données sur la Figure 1. A quoi correspondent les lignes pointillés sur la Figure 1? Comment les obtient-on?

Solution: Les lignes pointillées forment un intervalle de confiance autour des taux bruts. On sait que

$$\widehat{q_x} \sim \text{N}\left(q_x, \frac{q_x(1-q_x)}{E_x^0}\right).$$

On en déduit que l'intervalle de confiance de niveau α est donnée par

$$q_x \in \left[\widehat{q_x} - z_{\alpha/2}\sqrt{\frac{q_x(1-q_x)}{E_x^0}}, \widehat{q_x} + z_{\alpha/2}\sqrt{\frac{q_x(1-q_x)}{E_x^0}}\right].$$

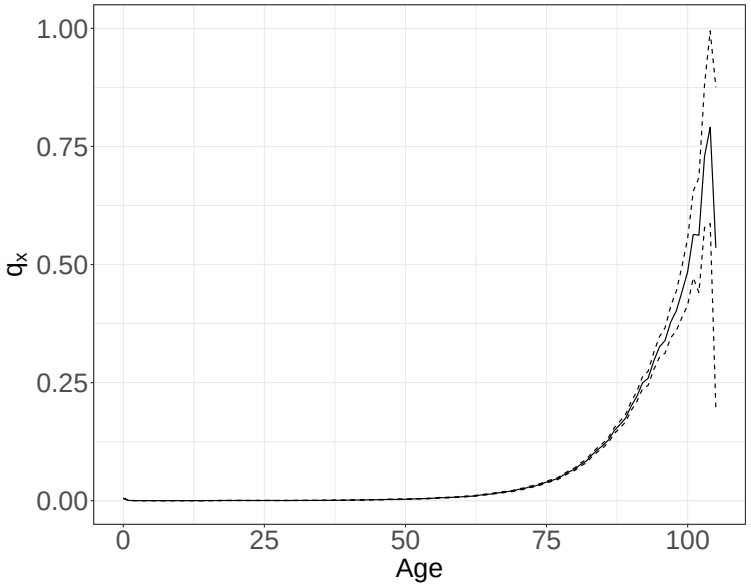

Figure 1: Probabilités de décès brutes

(c) (2 points) Quels sont les deux traitements à appliquer sur ces probabilités de décès pour obtenir une estimation finale des probabilité de décès jusqu'à l'âge 115? Vous devez détailler une méthode pour réaliser chacun des deux traitements.

Solution: Il faut appliquer une méthode de lissage etune méthode de fermeture de table. Les méthodes sont décrites dans le cours.

(d) (1 point) Donner une formule d'estimation de la fonction de survie $\widehat{S}(x)$ de X, la variable aléatoire égale à l'âge de décès. On donnera cette formule en fonction des $\widehat{q}_x$.

Solution: On peut estimer la fonction de survie par

$$\widehat{S}(x) = \prod_{k=0}^{x-1}(1-\widehat{q}_k), \text{ pour } x > 1,$$

et $S(0) = 1$. qui correspond à l'estimateur de Kaplan-Meier.

(e) (1 point) Un extrait de la table de mortalité issue des taux bruts est donné dans le Tableau 2. Comment est obtenue cette table ? A quoi correspond l_x?

Age	l_x
0	100,000
1	99,482
2	99,436
⋮	⋮
104	9
105	1
106	0

Table 2: Extrait de la table de mortalités au Pays-Bas pour l'année 2000.

Solution: On décide d'un radix, ici $l_0 = 100,000$ puis

$$l_x = l_0 \cdot \widehat{S}(x), \forall x.$$

(f) (1 point) Le modèle de Gompertz-Makeham suppose que la variable aléatoire X admet une fonction de hasard de la forme

$$h(x) = a + bc^x, \ x \in \mathbb{R}_+.$$

Soit $\theta = (a, b, c)$. Montrer que les probabilités de décès s'écrivent sous la forme

$$q_x(\theta) = 1 - sg^{c^x(c-1)},$$

où vous exprimerez s et g en fonction de a, b et c

Solution: On a

$$\begin{aligned} q_x(\theta) &= 1 - \exp\left(-\int_x^{x+1} a + bc^x \mathrm{d}x\right) \\ &= 1 - \exp\left(-a - b\int_x^{x+1} e^{x\log(c)}\mathrm{d}x\right) \\ &= 1 - e^a \exp\left(-\frac{b}{\log(c)}c^x(c-1)\right) \end{aligned}$$

On en déduit que

$$q_x(\theta) = 1 - sg^{c^x(c-1)},$$

avec $s = e^{-a}$ et $g = e^{-b/\log(c)}$.

(g) (1 point) Montrer que pour $q_x(\theta)$ proche de 0, on a

$$q_x(\theta) \approx -\log(s) - \log(g)(c-1)c^x.$$

Indication: On pourra utiliser un développement limité.

Solution: Pour $q_x(\theta)$ proche de 0, on a

$$q_x(\theta) \approx -\log(1 - q_x(\theta)) = -\log(s) - \log(g)(c-1)c^x.$$

(h) (2 points) En utilisant la question précédente identifier les coefficients α et β tel que

$$\log(q_{x+1}(\theta) - q_x(\theta)) \approx \alpha + \beta x$$

en fonction de θ. En déduire une méthode d'estimation pour g et c.

Solution: En utilisant l'approximation de la question précédente, il vient

$$\begin{aligned} \log(q_{x+1} - q_x) &\approx \log\left[\log(1/g)(c-1)^2c^x\right] \\ &= \log(\log(1/g)) + 2\log(c-1) + x\log(c) \\ &= \alpha + \beta x \end{aligned}$$

Cette relation suggère une regression linéaire de $\log(q_{x+1} - q_x)$ sur x et donc une estimation de α et β par les moindre carrés ordinaires.

(i) (1 point) En supposant connu g et c, comment estimer s?

Solution: On reprend l'aproximation précédente

$$q_x(\theta) \approx -\log(s) - \log(g)(c-1)c^x.$$

On peut donc écrire

$$s = \exp(-q_x - c^x(c-1)\log(g))$$

et prendre la moyenne pour tout x par exemple.

(j) (1 point) La Figure 2 montre les probabilités de décès issues de la table de mortalité 1 et du modèle de Makeham. Quels commentaires peut-on faire?

Solution: L'ajustement aux taux lissés n'est pas parfait. On constate une sur-estimation de la mortalité aux jeunes âges et une sous-estimation aux âges élevés.

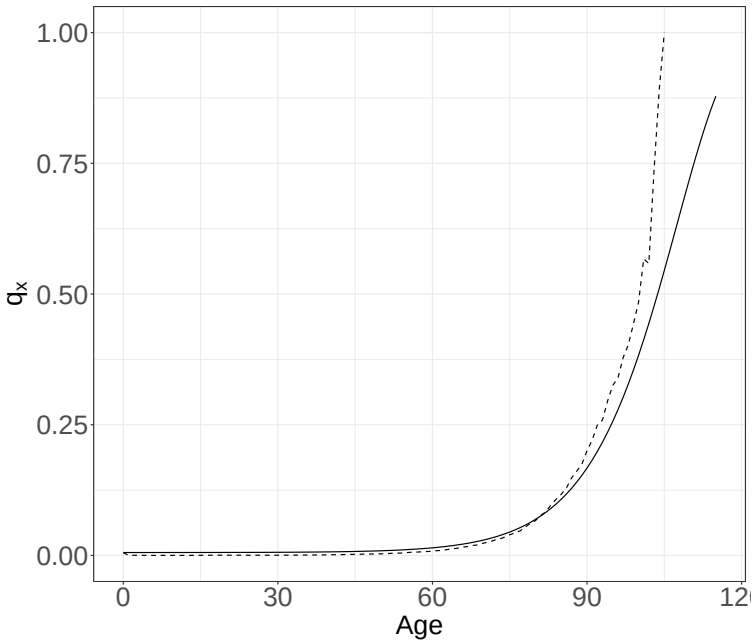

Figure 2: Probabilités de décès issues de la table de mortalité 2 (pointillé) et du modèle de Makeham (trait plein).

(k) (1 point) La Figure 3 montre les graphiques de $\log(q_{x+1} - q_x)$ et $\exp(-q_x - c^x(c-1)\log(g))$ en fonction de x. Comment pourrait-on améliorer l'ajustement du modèle de Makeham de la Figure 2?

Solution: Une solution serait d'introduire une pondération (w_x) des âges x, en supprimant par exemple les valeurs aberrantes visibles sur la figure 3.

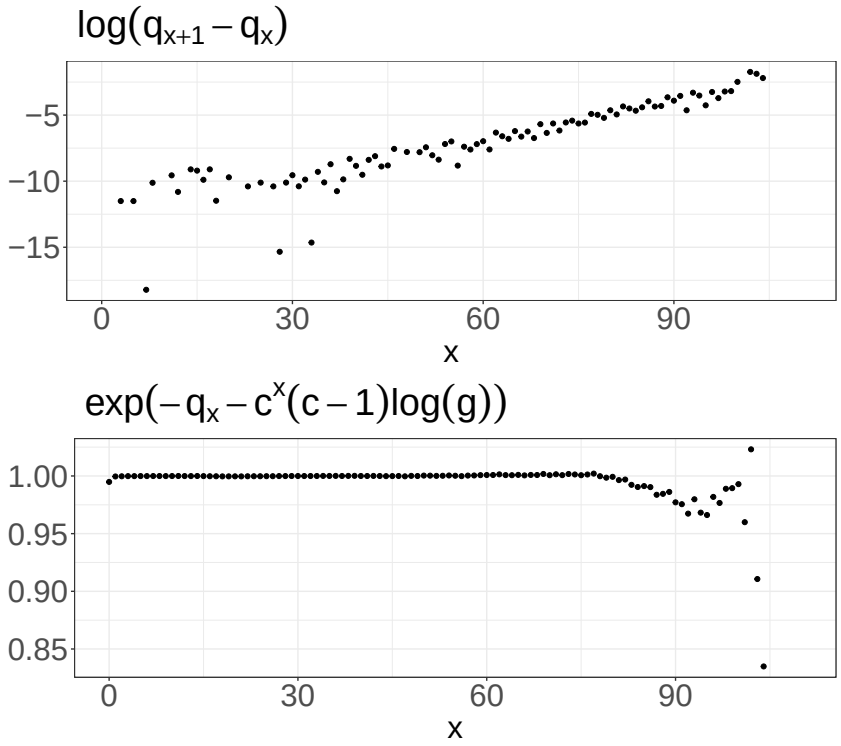

Figure 3: $\log(q_{x+1} - q_x)$ et $\exp(-q_x - c^x(c-1)\log(g))$ en fonction de x.