Université Claude Bernard - Lyon 1 Année 2015-2016
ISFA - M2SAF
Quentin Guibert
quentin.guibert@etu.univ-lyon1.fr

Travaux dirigés : modèles de durée
Séance n°1 du 25 Novembre 2015 - Corrigé

Exercice 1 Modèle de Gompertz-Makeham et fragilité Gamma.

Soit T la variable aléatoire positive représentant la durée de vie d'un individu que l'on suppose définie à partir du modèle de Gompertz-Makeham dont la fonction de hasard vaut $h(t) = a + bc^t$, avec a et b deux paramètres.

1. Rappeler pour ce modèle l'expression de sa fonction de survie, sa densité et sa fonction de survie conditionnelle à l'évènement $T > t_0$.
2. Écrire la log-vraisemblance du modèle et calculer sa dérivée première pour une échantillon $(t_1, \ldots, t_n)$
3. Pour une population donnée, on cherche à présent à mesurer l'effet d'une source d'hétérogénéité latente représentée par une variable Z, de densité $\pi(z)$ en date $t = 0$. 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. On suppose que la fragilité suit initialement une loi Gamma (λ, k) de densité

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

Donner l'expression de $\pi_t(z)$, la densité de la fragilité pour la population des survivants en date t, notée $Z(t)$. Fournir ensuite l'expression de la fonction de hasard moyenne pour la population $\bar{h}(t)$. En dérivant l'espérance de $Z(t)$ en date t, commenter et proposer un paramétrage possible permettant d'interpréter facilement ce modèle.

4. En revenant au cas général, écrire la vraisemblance du modèle avec fragilité, puis en intégrant selon Z en déduire l'expression de sa vraisemblance marginale (ou observable).

Réponse de l'exercice 1.

1. Rappel de cours :
   - fonction de survie : $S(t) = \exp\left(-\int_0^t a + bc^u du\right) = \exp\left(-at - \frac{b}{\ln(c)}\left(c^t - 1\right)\right)$;
   - densité : $f(t) = \left(a + bc^t\right)\exp\left(-at - \frac{b}{\ln(c)}\left(c^t - 1\right)\right)$;
   - fonction de survie conditionnelle : $\frac{S(t+t_0)}{S(t_0)} = \exp\left(-at - \frac{bc^{t_0}}{\ln(c)}\left(c^t - 1\right)\right)$.

Dans la suite, on note $H(t) = \int_0^t h(u)\,du$ la fonction de hasard cumulée.

La loi de Gompertz-Makeham est usuellement utilisée pour modéliser la durée de vie humaine. Le second terme du taux de hasard (bc^t) correspond à la loi de Gompertz originale (1825) et est croissant pour $c > 1$ et $b \geq 0$. Il permet de traduire le vieillissement progressif de l'organisme. Le paramètre a ajouté par Makeham intègre les décès accidentels survenant aux âges plus jeunes. Si ce modèle est retenu pour la population générale, il ne permettra pas de prendre en compte les âges de la vie où le taux de hasard est potentiellement décroissant (mortalité infantile, bosses dues aux accidents chez les jeunes adultes...). Notons qu'il n'existe pas d'expression explicite pour la loi de Makeham-Gompertz pour l'espérance de vie et les moments d'ordre supérieurs.

2. La log-vraisemblance du modèle s'écrit

$$\ln\left(\mathcal{L}(a,b,c)\right) = \sum_i^n \ln\left(a + bc^{t_i}\right) - at_i - \frac{b}{\ln(c)}\left(c^{t_i} - 1\right).$$

Ainsi, ses dérivées première valent

$$\frac{\partial \ln\left(\mathcal{L}(a,b,c)\right)}{\partial a} = \sum_{i=1}^n \frac{1}{a + bc^{t_i}} - t_i \ ,$$

$$\frac{\partial \ln\left(\mathcal{L}(a,b,c)\right)}{\partial b} = \sum_{i=1}^n \frac{c^{t_i}}{a + bc^{t_i}} - \frac{c^{t_i} - 1}{\ln(c)} \ ,$$

$$\frac{\partial \ln\left(\mathcal{L}(a,b,c)\right)}{\partial c} = \sum_{i=1}^n \frac{t_i b c^{t_i - 1}}{a + bc^{t_i}} - \frac{t_i b c^{t_i - 1}}{\ln(c)} + \frac{b\left(c^{t_i} - 1\right)}{c\left(\ln(c)\right)^2}.$$

L'estimateur de maximum de vraisemblance satisfait le système où ces trois équations sont nulles. Le système se résout numériquement par le biais d'un procédure itérative (ex : algorithme de Newton-Raphson).

3. Avec le temps, les proportions d'individus prenant une même valeur pour Z se modifie. Ainsi, on a pour la population des survivants en date t, i.e. sachant $T \geq t$

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

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

Or

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

avec $\lambda(t) = \lambda + H(t)$. Le terme $\dfrac{\lambda^k}{\left(\lambda(t)\right)^k}$ se simplifie au numérateur et au dénominateur de $\pi_t(z)$ et on reconnait une loi Gamma de paramètre $(k, \lambda(t))$.

La fonction de hasard moyenne s'écrit

$$\begin{aligned}\bar{h}(t) &= \int_0^\infty h(t|z)\,\pi_t(z)\,dz = h(t)\int_0^\infty z\pi_t(z)\,dz = h(t)\,\frac{k}{\lambda(t)} \\ &= \frac{k\left(a+bc^t\right)}{\lambda + at + \frac{b}{\ln(c)}\left(c^t-1\right)}.\end{aligned}$$

L'espérance de $Z(t)$ en t s'écrit

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

On a alors

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

avec $\sigma(t)^2$ la variance de $Z(t)$. Tout d'abord, on observe que la moyenne de la fragilité décroit avec le temps, puisque les décès conduisent à toucher plus tôt les individus pour lesquels la fragilité est la plus grande.
En supposant que $\lambda = k = \frac{1}{\sigma^2}$, on obtient

$$\bar{h}(t) = \frac{a+bc^t}{1+\sigma^2\left(at + \frac{b}{\ln(c)}\left(c^t-1\right)\right)}.$$

On note également que h croît plus vite que $\bar{h}$ et qu'il est possible d'obtenir

$$\bar{h}(t) = h(t)\,S()^{\sigma^2}.$$

4. La vraisemblance (complète) du modèle s'écrit

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^n \left(zh(t_i|z)\exp\left(-zH(t_i)\right)\right)\pi(z) \\ &= \prod_{i=1}^n \left(zh(t_i|z)\exp\left(-zH(t_i)\right)\right)\frac{\lambda^k z^{k-1}\exp(-\lambda z)}{\Gamma(k)}\ ,\end{aligned}$$

avec $\boldsymbol{\theta} = (a,b,c,\lambda,k)$. De manière analogue à ce qui a été fait précédemment, on remarque qu'il est possible de faire apparaître la densité d'une loi de Gamma. On introduit pour cela la notation $\mu = \lambda + \sum_{i=1}^n H(t_i)$ et $\nu = k+n$

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{\lambda^k\Gamma(\nu)\prod_{i=1}^n h(t_i)}{\Gamma(k)\,\mu^\nu}\frac{\mu^\nu z^{\nu-1}\exp(-\mu z)}{\Gamma(\nu)}.$$

L'inférence du modèle peut être simplement obtenu par maximum de vraisemblance. On considère la vraisemblance marginale en intégrant selon z, ce qui donne immédiatement

$$\mathcal{L}_{\text{marg}}(\boldsymbol{\theta}) = \frac{\lambda^k\Gamma(\nu)\prod_{i=1}^n h(t_i)}{\Gamma(k)\,\mu^\nu}\ ,$$

et la log-vraisemblance est

$$\ln\mathcal{L}_{\text{marg}}(\boldsymbol{\theta}) = k\ln(\lambda) + \ln(\Gamma(\nu)) + \sum_{i=1}^n \ln(h(t_i)) - \ln(\Gamma(k)) - \nu\ln(\mu).$$

Dériver cette quantité pour obtenir le score de ce modèle ne pose pas de difficulté particulière.

Exercice 2 Vraisemblance et données manquantes.

Dans la plupart des analyse de survie, les observations comprennent des données manquantes. Soit un échantillon de n individus de durée de vie respective $T_1, \ldots, T_n$ i.i.d. Ces observations suivent une loi de densité f définie à partir du paramètre $\boldsymbol{\theta}$.

1. Dans chacun des cas suivant, écrire l'expression générale de la vraisemblance, puis introduire progressivement (a) l'hypothèse d'indépendance du processus lié aux données manquantes et (b) le fait que ce processus ne dépende pas de $\boldsymbol{\theta}$ (non-informatif).
   - censure C à droite de type I ;
   - troncature individuelle à gauche L_i ;
   - censure individuelle à gauche L_i et à droite R_i.
2. On suppose à présent que les observations sont uniquement soumises à censure **indépendante** à droite C_i. On fait l'hypothèse que les T_i et les C_i ont pour fonction de hasard respective $h_T(t) = \alpha t^{\alpha-1}$ et $h_C(t) = \beta t^{\beta-1}$ (Weibull). Écrire la log-vraisemblance du modèle et en déduire l'équation vérifie par l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta} = (\alpha, \beta)$.

Réponse de l'exercice 2.

1. **Censure type I**
   On observe (Y_i, D_i) pour chaque i, où $Y_i = T_i \wedge C$ et $D_i = \mathbb{1}_{\{Y_i = T_i\}}$. La contribution à la vraisemblance individuelle s'écrit

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}) &= \mathbb{P}(Y_i = y_i, D_i = d_i; \boldsymbol{\theta}) \\ &= \mathbb{P}(T_i = y_i, D_i = 1; \boldsymbol{\theta})^{d_i} \mathbb{P}(C = y_i, D_i = 0; \boldsymbol{\theta})^{1-d_i} \\ &= \mathbb{P}(T_i = y_i, T_i \leq C; \boldsymbol{\theta})^{d_i} \mathbb{P}(C = y_i, C \leq T_i; \boldsymbol{\theta})^{1-d_i}.\end{aligned}$$

   Si la censure est indépendante, on a

$$\mathcal{L}_i(\boldsymbol{\theta}) = (f_T(y_i; \boldsymbol{\theta}) S_C(y_i; \boldsymbol{\theta}))^{d_i} (f_C(y_i; \boldsymbol{\theta}) S_T(y_i; \boldsymbol{\theta}))^{1-d_i}.$$

   Si la censure est non-informative, on se limite à

$$\mathcal{L}_i(\boldsymbol{\theta}) \propto (f_T(y_i; \boldsymbol{\theta}))^{d_i} (S_T(y_i; \boldsymbol{\theta}))^{1-d_i}.$$

   Troncature à gauche
   On observe (L_i, T_i) pour chaque i si $L_i \leq T_i$. La contribution à la vraisemblance individuelle s'écrit

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}) &= \mathbb{P}(T_i = t_i L_i = l_i | L_i \leq T_i; \boldsymbol{\theta}) \\ &= \frac{1}{S_T(l_i; \boldsymbol{\theta})} f_{(T,L)}(t_i, l_i; \boldsymbol{\theta}) \mathbb{1}_{l_i \leq t_i}.\end{aligned}$$

Si la troncature est indépendante et comme on n'observe que des durées au-delà de la date de troncature, on a

$$\mathcal{L}_i(\boldsymbol{\theta}) = \frac{f_T(t_i;\boldsymbol{\theta})\, f_L(l_i;\boldsymbol{\theta})}{S_T(l_i;\boldsymbol{\theta})}.$$

Si la troncature est non-informative, on se limite à

$$\mathcal{L}_i(\boldsymbol{\theta}) = \frac{f_T(t_i;\boldsymbol{\theta})}{S_T(l_i;\boldsymbol{\theta})}.$$

Remarque : on voit clairement qu'en présence de censure à droite et de troncature à gauche indépendantes et non-informatives, on aurait

$$\mathcal{L}_i(\boldsymbol{\theta}) \propto \frac{1}{S_T(l_i;\boldsymbol{\theta})} \left(f_T(t_i;\boldsymbol{\theta})\right)^{d_i} \left(S_T(t_i;\boldsymbol{\theta})\right)^{1-d_i}.$$

Censure individuelle à gauche et à droite

On introduit $(\Delta_{1,i}, \Delta_{2,i}, \Delta_{3,i})$ pour chaque i, où $\Delta_{1,i} = \mathbb{1}_{\{T_i \leq L_i\}}$, $D_{2,i} = \mathbb{1}_{\{L_i < T_i \leq R_i\}}$ et $D_{3,i} = \mathbb{1}_{\{R_i < T_i\}}$. On note $Y_i = (T_i \wedge R_i) \vee L_i$. La contribution à la vraisemblance individuelle s'écrit

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}) &= \mathbb{P}(L_i = y_i, \Delta_{1,i} = 1;\boldsymbol{\theta})^{\delta_{1,i}}\, \mathbb{P}(T_i = y_i, \Delta_{2,i} = 1;\boldsymbol{\theta})^{\delta_{2,i}}\, \mathbb{P}(R_i = y_i, \Delta_{3,i} = 1;\boldsymbol{\theta})^{\delta_{3,i}} \\ &= \mathbb{P}(L_i = y_i, T_i \leq L_i;\boldsymbol{\theta})^{\delta_{1,i}}\, \mathbb{P}(T_i = y_i, L_i < T_i \leq R_i;\boldsymbol{\theta})^{\delta_{2,i}}\, \mathbb{P}(R_i = y_i, R_i < T_i;\boldsymbol{\theta})^{\delta_{3,i}}.\end{aligned}$$

Si la censure est indépendante, i.e. $T \perp\!\!\!\perp (L, R)$, on a

$$\mathcal{L}_i(\boldsymbol{\theta}) = \left(f_L(y_i;\boldsymbol{\theta})\, F_T(y_i;\boldsymbol{\theta})\right)^{\delta_{1,i}} \left(f_T(y_i;\boldsymbol{\theta})\, \mathbb{P}(L_i < y_i \leq R_i;\boldsymbol{\theta})\right)^{\delta_{2,i}} \left(f_R(y_i;\boldsymbol{\theta})\, S_T(y_i;\boldsymbol{\theta})\right)^{\delta_{3,i}}.$$

Si la censure est non-informative, on se limite à

$$\mathcal{L}_i(\boldsymbol{\theta}) \propto F_T(y_i;\boldsymbol{\theta})^{\delta_{1,i}}\, f_T(y_i;\boldsymbol{\theta})^{\delta_{2,i}}\, S_T(y_i;\boldsymbol{\theta})^{\delta_{3,i}}.$$

2. Si la censure est indépendante, la contribution individuelle à la vraisemblance s'écrit

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}) &= \left(f_T(y_i;\alpha)\, S_C(y_i;\beta)\right)^{d_i} \left(f_C(y_i;\beta)\, S_T(y_i;\alpha)\right)^{1-d_i} \\ &= \left(\alpha y_i^{\alpha-1}\right)^{d_i} \left(\beta y_i^{\beta-1}\right)^{1-d_i} \exp\left(-y_i^{\alpha}\right) \exp\left(-y_i^{\beta}\right).\end{aligned}$$

La log-vraisemblance de l'échantillon est donc

$$\ln \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} d_i \left(\ln(\alpha) + (\alpha-1)\ln(y_i)\right) + (1-d_i)\left(\ln(\beta) + (\beta-1)\ln(y_i)\right) - y_i^{\alpha} - y_i^{\beta}.$$

L'EMV de $\boldsymbol{\theta}$ s'obtient numériquement en résolvant les deux équations

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta})}{\partial \alpha} = \sum_{i=1}^{n} \frac{d_i}{\alpha} + d_i \ln(y_i) - y_i^{\alpha} \ln(y_i) = 0 \ ,$$

et

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta})}{\partial \beta} = \sum_{i=1}^{n} \frac{1-d_i}{\beta} + (1-d_i) \ln(y_i) - y_i^{\beta} \ln(y_i) = 0.$$

Bonus : Quelques exemples de censure et de troncature.

- censure à droite : les sorties d'une étude épidémiologique, la résiliation d'un contrat d'assurance → l'évènement d'intérêt (par exemple le décès) se produit après la perte d'information ;
- censure à gauche : observation du temps de descente des babouins de leur arbre (tiré de la littérature)→ censure à gauche si le babouin est descendu avant l'arrivée de l'observateur ;
- troncature à gauche : pour un contrat d'assurance vie, on dispose de la date de souscription (troncature) et éventuellement de la date de décès (censure à droite). Cet échantillon n'est pas représentatif de la population et ne donne pas d'information sur la durée de vie avant la date de troncature.

Exercice 3 Estimation d'un modèle à hasard constant par morceaux.

Soit un échantillon de n individus i.i.d. de durée de vie $T_1, \ldots, T_n$. Chaque observation i est soumise à censure à droite C_i, supposée indépendante et non-informative. Pour une décomposition en K segment fixés de l'espace de valeurs prises par la durée de vie, la fonction de hasard du modèle est supposée constante par morceaux (appelé aussi modèle de Poisson) telle que

$$h(t) = \theta_k \text{ pour } t \in J_k =]t_{k-1}; t_k] \ , \ k = 1, \ldots, K,$$

où $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ est un vecteur de paramètres à estimer.

1. Écrire la log-vraisemblance du modèle censuré, calculer son score 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.
2. Calculer la matrice d'information de Fisher. En déduire une expression de la variance asymptotique des $\widehat{\theta}_k$ et proposer un intervalle de confiance asymptotique de niveau α pour ces estimateurs. Préciser quel serait l'estimateur retenu en cas de troncature à gauche (indépendante et non-informative)
3. Expliciter un test pour l'hypothèse $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.
4. On considère à présent un modèle de régression pour la durée de vie humaine pour lequel les fonctions de hasard sont supposées constantes par chaque âge entier $x \in \{x_{\min}, \ldots, x_{\max}\}$. Le modèle se présente sous la forme d'un modèle à hasard proportionnel tel que

   $$h(t) = \exp(\eta(t)) \ ,$$

   où pour $t \in J_x =]x; x+1] \ , \ x = x_{\min}, \ldots, x_{\max}$, le prédicteur linéaire vaut

   $$\eta(t) = \sum_{s=0}^{p} \beta_s x^s \ ,$$

   avec $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ le vecteur de paramètres à estimer.
   Écrire la vraisemblance de ce modèle en présence de censure à droite indépendante, puis le système permettant d'obtenir l'estimateur du maximum de vraisemblance.

5. Proposer comment contrôler la qualité de l'ajustement réalisé.

Réponse de l'exercice 3.

1. En présence de censure à droite indépendante et non-informative, on observe (Y_i, D_i) pour chaque i, où $Y_i = T_i \wedge C_i$ et $D_i = \mathbb{1}_{\{T_i \leq C_i\}}$. La log-vraisemblance s'écrit

$$\ln\left(\mathcal{L}\left(\boldsymbol{\theta}\right)\right) = \sum_{i=1}^{n} d_i \ln\left(h\left(y_i\right)\right) - \int_0^{y_i} h\left(u\right) du.$$

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

$$d_i \ln\left(h\left(y_i\right)\right) = \sum_{k=1}^{K} d_{i,k} \ln\left(\theta_k\right) ,$$

et

$$\int_0^{y_i} h\left(u\right) du = \sum_{k=1}^{K} \left(t_k \wedge y_i - t_{k-1}\right)_+ \theta_k = \sum_{k=1}^{K} y_{ik}\theta_k ,$$

avec y_{ik} le temps passé par l'individu i dans le segment k. Ainsi, on en déduit

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

avec N_k le nombre de décès observés dans le segment k et R_k 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

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

On parle donc usuellement de modèle de Poisson pour le nombre de décès.
En dérivant par rapport à θ_k, la $k-$ième composante du vecteur de score s'écrit

$$U_k\left(\theta\right) = \frac{N_k}{\theta_k} - R_k ,$$

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.

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

$$\mathcal{I}_{kl}\left(\boldsymbol{\theta}\right) = \delta_{kl}\frac{N_k}{\theta_k^2} ,$$

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

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

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

3. Trois tests classiques peuvent être proposés pour tester l'hypothèse $\boldsymbol{\theta} = \boldsymbol{\theta}_0$:
   – statistique de Wald : $\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^\top \mathcal{I}\left(\hat{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$;
   – statistique de score : $\boldsymbol{U}\left(\boldsymbol{\theta}_0\right)^\top \mathcal{I}\left(\boldsymbol{\theta}_0\right)^{-1} \boldsymbol{U}\left(\boldsymbol{\theta}_0\right)$;
   – ratio des vraisemblance : $2\left(\ln \mathcal{L}\left(\hat{\boldsymbol{\theta}}\right) - \ln \mathcal{L}\left(\boldsymbol{\theta}_0\right)\right)$.
   Explicitons par exemple la statistique de Wald

$$\begin{aligned}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^\top \mathcal{I}\left(\hat{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) &= \sum_{k=1}^{K}\left(\hat{\theta}_k - \theta_0\right)^2 \mathcal{J}_{kk}\left(\hat{\boldsymbol{\theta}}\right) \\ &= \sum_{k=1}^{K}\left(\frac{N_k}{R_k} - \theta_0\right)^2 \frac{R_k^2}{N_k}.\end{aligned}$$

4. En utilisant des notations évidentes, la vraisemblance du modèle s'écrit en regroupant par âge entier

$$\ln\left(\mathcal{L}\left(\boldsymbol{\beta}\right)\right) = \sum_{x=x_{\min}}^{x_{\max}} N_x \eta\left(x\right) - \sum_{x=x_{\min}}^{x_{\max}} R_x \exp\left(\eta\left(x\right)\right).$$

   On remarque que considérer un modèle linéaire généralisé de Poisson (lien canonique log) tel que

$$N_x \sim \mathcal{P}\left(R_x \exp\left(\eta\left(x\right)\right)\right).$$

   Il est donc possible de se placer dans le cadre de ce type de modèle de régression pour la suite du problème.
   Le système permettant d'obtenir les paramètres du modèle est

$$\frac{\partial}{\partial \beta_0} \ln\left(\mathcal{L}\left(\boldsymbol{\beta}\right)\right) = \sum_{x=x_{\min}}^{x_{\max}} N_x - R_x \exp\left(\eta\left(x\right)\right) = 0\ ,$$

   et pour tout $s = 1, \ldots, p$,

$$\frac{\partial}{\partial \beta_s} \ln\left(\mathcal{L}\left(\boldsymbol{\beta}\right)\right) = \sum_{x=x_{\min}}^{x_{\max}} x^s \left(N_x - R_x \exp\left(\eta\left(x\right)\right)\right) = 0.$$

   La première relation permet d'assurer la reproduction du nombre total de décès par le modèle. En posant le résidu

$$r_x = N_x - R_x \exp\left(\eta\left(x\right)\right)\ ,$$

   les équations suivantes apparaissent comme une relation d'orthogonalité entre les résidus et variables x^s. Ces équations se résolvent de la même manière que pour un modèle linéaire généralisé de Poisson.
5. Comme pour un modèle linéaire généralisé classique, la qualité du modèle peut être analysée par le biais de :
   – la déviance

$$\begin{aligned}\mathrm{Dev}\left(\boldsymbol{N}, \hat{\boldsymbol{N}}\right) &= 2\left(\ln\left(\mathcal{L}\left(\hat{N}_x\right)\right) - \ln\left(\mathcal{L}\left(N_x\right)\right)\right) \\ &= 2\sum_{x=x_{\min}}^{x_{\max}} N_x \ln\left(\frac{N_x}{R_x \exp\left(\hat{\eta}\left(x\right)\right)}\right) - \left(N_x - R_x \exp\left(\hat{\eta}\ \ (x)\right)\right) \\ &= 2\sum_{x=x_{\min}}^{x_{\max}} N_x \ln\left(\frac{N_x}{R_x \exp\left(\hat{\eta}\left(x\right)\right)}\right)\end{aligned}$$

   garanti par la présence de β_0.

– test d'adéquation du chi-2 de Pearson

$$\chi^2 = \sum_{x=x_{\min}}^{x_{\max}} \frac{\left(N_x - R_x \exp\left(\hat{\eta}\ (x)\right)\right)^2}{R_x \exp\left(\hat{\eta}\ (x)\right)}$$

– le pseudo-R^2

$$\text{pseudo-}R^2 = 1 - \frac{\text{Dev}\left(\boldsymbol{N}, \hat{\boldsymbol{N}}\right)}{\text{Dev}\left(\boldsymbol{N}, \bar{\boldsymbol{N}}\right)}.$$

Il convient également d'analyser les résidus et l'influence des points.

Exercice 4 Modèles AFT et avec *odds* proportionnels.

On souhaite étudier deux classes de modèles classiques de régression : les modèles *accelered 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 $\boldsymbol{X}$ un vecteur de covariables.

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

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

   avec $\boldsymbol{\theta}$ un vecteur de paramètres et une distribution correspondant à un terme d'erreur. Écrire la loi de survie et la fonction de hasard du modèle en fonction de celles de la loi de référence $T_0 = \exp(W)$.
2. On s'intéresse à présent au second type de modèle que l'on suppose défini par la relation

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

   avec $S_0^*(t)$ la fonction de survie d'une loi quelconque et $\boldsymbol{\beta}$ un vecteur de paramètres. Montrer dans le cas particulier où T_0 suit une loi de de Weibull

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

   que la loi du modèle AFT est encore une loi de Weibull, mais que dans ces conditions elle ne vérifie pas l'hypothèse PO.
3. Caractériser la loi de S_0 pour que le modèle AFT vérifie l'hypothèse PO. En prenant différentes valeurs de $\boldsymbol{X}$, on cherchera à maintenir le ratio $\dfrac{\beta_j}{\theta_j}$ constant.

Réponse de l'exercice 4.

1. En passant à l'exponentielle, on a

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

   D'où l'on tire immédiatement

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

   et

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

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

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

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

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

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

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

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

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

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

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

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

3. 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

$$S_0^*\left(t\right) = S_0\left(t\right).$$

En prenant $\boldsymbol{X} = \left(-\dfrac{1}{\theta_1}\ln\left(t\right),0,0,\ldots\right)$, l'égalité devient

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

et donc pour tout $t \geq 0$

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

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

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

D'où

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

et donc

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

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