Modèles de survie hétérogènes

Randy Laine et Robin Ryder
sous la direction de François Taddei et Grégory Paul

Juin 2005

Table des matières

1 Présentation du problème

1.1 C. Elegans : un nématode intéressant

Peu encombrant, facile à cultiver en laboratoire, inoffensif pour l'homme et l'environnement (on en trouve de grandes quantités dans la terre), ce petit ver cylindrique d'environ un millimètre de long présente de nombreuses autres caractéristiques qui en font un excellent sujet d'étude pour les biologistes. À mi-chemin entre la bactérie unicellulaire et un organisme complexe de plusieurs centimètres, le développement de Caenorhabditis Elegans est très bien connu et se fait avec une précision assez remarquable. Tout individu normal présente exactement 959 cellules différenciées en cellules nerveuses, musculaires, intestinales et épidermiques. Ces cellules étant assez transparentes (les cellules de peau ne cachent pas celles qui sont en dessous), on peut les identifier individuellement et suivre leurs divisions, migrations et réorganisation qui se déroulent d'ailleurs toujours selon la même procédure. En outre, C. Elegans est principalement hermaphrodite (il existe aussi des mâles mais en très faible proportion) d'où une reproduction par clonage qui permet d'obtenir assez rapidement un grand nombre de vers tous génétiquement identiques. Ce ver possède une dernière propriété qui nous intéresse : il passe par des stades de développement que l'on retrouve chez tous les animaux plus complexes : division et différenciation cellulaire à partir d'une seule cellule, naissance, croissance ("enfance et adolescence", dirait-on chez les humains), reproduction, vieillissement et dégénérescence des fonctions (fonctions motrices ...) et finalement, décès.

C. Elegans est donc une sorte de "modèle simplifié" dont l'étude est une étape vers une meilleure compréhension des organismes plus complexes. On peut par ailleurs facilement introduire de nouveaux gènes ou détruire certaines cellules par un faisceau laser afin de tester différentes hypothèses. D'autres expérimentations et observations plus "douces" sont aussi intéressantes, comme celle qui a été mise en place au laboratoire TaMaRa (INSERM U571) dont les mesures expérimentales servent de données pour cet exposé.

FIG. 1 – *C. Elegans*

1.2 Description et enjeu de l'expérience

Des vers C. Elegans sont cultivés dans des boîtes de Petri dont le fond est recouvert d'un tapis nutritif sur lequel sont déposées des bactéries. Les bactéries se nourrissent du tapis nutritif et sont mangées par C. Elegans.

Les données que nous exploitons ont été obtenues en utilisant trois souches de bactéries différentes : OP50, F11 et ED1A (au sein d'une même souche, les bactéries sont génétiquement identiques), ainsi que des vers C. Elegans tous génétiquement identiques. Dans chaque boîte de Petri ont été introduits une cinquantaine de vers et une seule souche de bactéries, et on note chaque jour le nombre de vers qui meurent. Cette expérience est alors répétée 3 fois, et nous disposons donc de 9 séries de mesures (3 séries pour chaque souche de bactérie). Fait troublant, la durée de vie des vers varie de 4 à 20 jours ! Quelle est l'origine d'une telle variabilité ? Il semble raisonnable au premier abord que c'est le type de bactérie mis en contact avec les vers qui conditionne la durée de vie moyenne et/ou maximale des vers, tout comme une alimentation saine et équilibrée est *a priori* un gage de meilleure santé chez les humains. Ceci semble d'autant plus vrai dans le cas des vers que les bactéries, introduites dans le système digestif des vers, interagissent en retour sur ces derniers. On peut alors émettre l'hypothèse que certaines bactéries sont plus "agressives" que d'autres, tout comme certains aliments sont supposés plus cancérigènes que d'autres si l'on désire poursuivre le parallèle avec les humains.

Ceci permet l'étude des bactéries pathogènes : leur virulence, mode d'agression (affaiblissement progressif continu ou bien par étapes ? Ou bien attaque foudroyante ? Cibles privilégiées : les jeunes ou les adultes ?) Par exemple l'objet d'une thèse a été l'étude d'uropathogènes. La procédure habituelle est de prélever des germes et de l'injecter à des souris que l'on place en observation. La difficulté est que le cycle de vie d'une souris est de 2 années, d'où l'idée d'utiliser des vers C. Elegans, bien connus et dont le cycle de vie inférieur à 20 jours. La première partie de l'étude a donc été de vérifier que l'effet des uropathogènes sur les souris se retrouvait au moins qualitativement chez les vers.

1.3 Les différents modèles

Le but de cet exposé est de déterminer les paramètres importantes de la loi $\lambda_i(t)$ que suit le taux de mortalité des vers. Pour ce faire, il existe traditionnellement trois approches.

Définition 1.3.1 (Approche paramétrique) *On fixe les paramètres $X_1, \ldots, X_k$ (en nombre fini et supposés connus) dont dépend la loi et en outre, on fait l'hypothèse que la loi de probabilité appartient à une classe particulière de fonctions de ces paramètres. Par exemple, on utilise souvent une loi normale si le support des variables est $\mathbb{R}$, et dans notre cas, où le support est dans $\mathbb{R}_+$, on préférera une*

distribution exponentielle. L'enjeu est donc de déterminer grâce à une méthode de régression et à partir des données expérimentales les coefficients $a_1, \ldots, a_k$. L'avantage de l'approche paramétrique est que la phase d'estimation des coefficients à partir des données est assez simple, mais nous ne la retiendrons pas ici car elle diffère souvent trop des données expérimentales.

Définition 1.3.2 (Approche non-paramétrique) *Cette fois-ci, on ne fait aucune hypothèse sur la forme de la loi de probabilité et donc, il n'y a pas besoin de recenser tous les paramètres importants. L'enjeu est alors d'estimer, d'interpoler la loi grâce aux données expérimentales. L'inconvénient de cette méthode est qu'elle nécessite beaucoup de données pour espérer obtenir une estimation fiable.*

Définition 1.3.3 (Approche semi-paramétrique) *Elle est un compromis entre les deux approches exposées précédemment : nous fixons les paramètres qui semblent importants tout en admettant que d'autres paramètres importants nous échappent. La loi de probabilité se compose alors d'un produit de deux fonctions f et g ; la première étant une fonction particulière des paramètres que l'on a choisis, et il faudra déterminer les coefficients par régression (approche plutôt paramétrique) ; la seconde est une fonction qu'il faudra déterminer par interpolation (approche plutôt non-paramétrique). L'intérêt de cette méthode est qu'elle permet d'espérer de trouver $\lambda(t)$ même si nous ne connaissons pas tous les paramètres importants ou si nous n'arrivons pas à obtenir des données. Cette approche est très utilisée en analyse de survie, surtout au travers du modèle de Cox, que nous avons choisi d'utiliser.*

1.4 Importance de l'hétérogénéité

Lors d'une étude, les expérimentateurs ne peuvent observer que la valeur moyenne de ce qu'ils observent; par exemple, le taux de survie ou la durée moyenne de vie sont par leurs définitions même des valeurs moyennes sur toute une population. On a alors tendance à interpréter le résultat observé comme pouvant s'appliquer à chaque individu séparément. L'hypothèse implicite est donc que la population est homogène. Par exemple, l'espérance de vie à la naissance de chaque individu est la même, et elle est égale à la durée de vie moyenne.

Cette hypothèse est clairement fausse : tous les individus n'ont pas les mêmes caractéristiques, et n'ont donc pas la même espérance de vie. On pourrait être tenté de penser que la moyenne (observée) reflète néanmoins assez bien la réalité, et que si le taux de mortalité moyen augmente, c'est que le taux de mortalité individuel a en moyenne augmenté.

Dans cette partie, nous exposerons quelques exemples montrant que la valeur moyenne observée peut être trompeuse quand il y a hétérogénéité. Nous étudierons des cas simples, où la population est divisée en deux groupes homogènes. Naturellement, ces exemples peuvent s'étendre à des cas où l'hétérogénéité

est plus importante, et qui refléteront souvent mieux la réalité.

On considère une population de taille $N(t)$ à l'instant t, et $p(t) = \frac{N(t)}{N(0)}$ le taux de survie. On s'intéresse à $p(t)$ et au taux de mortalité $\lambda(t)$, qui présentent l'avantage de ne pas dépendre de la taille de la population étudiée. On se place dans un monde où la ressurrection est impossible, les fonctions N et p sont donc décroissantes. L'équation d'évolution de la population est

$$N(t + dt) - N(t) = -\lambda(t)\, dt\, N(t)$$

Ou encore, en divisant par $N(0)$,

$$p(t + dt) - p(t) = -\lambda(t)\, dt\, p(t)$$

(Cette équation intuitive est en fait une définition possible du taux de mortalité.) D'où

$$p(t) = \exp\left[-\int_0^t \lambda(x)\, dx\right]$$

Notons qu'on peut remplacer la naissance et la mort par d'autres événements, par exemple la sortie de prison et la récidive, l'obtention d'un diplôme et la publication du premier article scientifique, l'abandon et la reprise de la cigarette, le mariage et le divorce, l'entrée et la sortie du chômage, etc.

On suppose que la population est divisée en deux sous-populations. Le point important est que l'expérimentateur ne sait pas comment ces deux sous-populations sont réparties, ni à quel groupe un individu donné appartient. Nous allons donc comparer la valeur observée (la moyenne) à la valeur réelle de chaque sous-population. En notant p_1, p_2, λ_1, λ_2 les taux de survie et de mortalité respectifs des deux sous-populations, et $\pi(t)$ la proportion de la sous-population 1 à l'instant t, on a

$$\pi(t) = \frac{\pi(0)\, p_1(t)}{\pi(0)\, p_1(t) + (1 - \pi(0))\, p_2(t)}$$

Le taux de mortalité observé est alors $\overline{\lambda} = \pi(t)\, \lambda_1(t) + (1 - \pi(t))\, \lambda_2(t)$

Dans la figure 2, on observe une mortalité décroissante au cours du temps. C'est ce qu'on observe notamment chez les anciens fumeurs : plus le temps passe, moins les anciens fumeurs se remettent à fumer. Cela ne veut pas forcément dire qu'un ancien fumeur donné a de moins en moins de chances de reprendre la cigarette. La même courbe peut être obtenue s'il y a deux types d'anciens fumeurs, les incurables et les guéris, chaque groupe ayant un taux de rechute constant ($\lambda_1 = 0.08; \lambda_2 = 0.02$). Avec le temps, de plus en plus d'incurables rechutent, la proportion d'incurables baisse, et le taux de rechute observé diminue. Mais

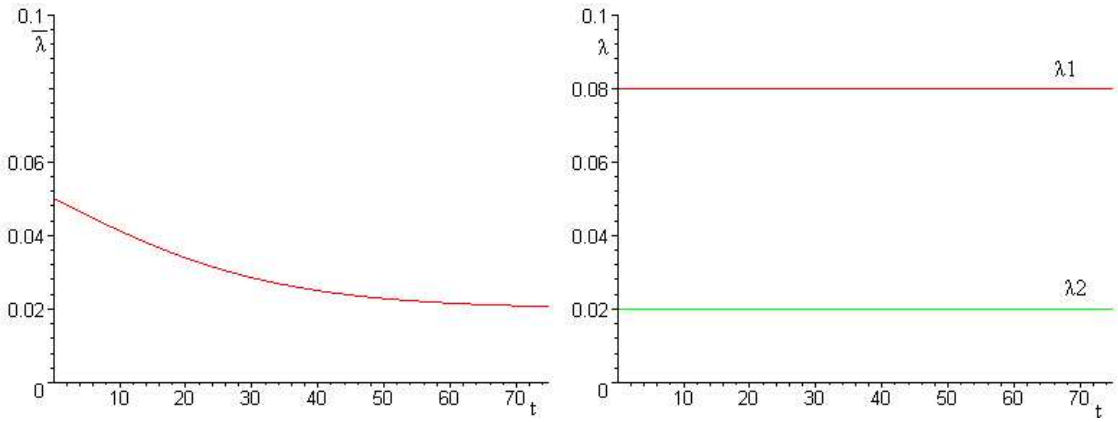

FIG. 2 – *La mortalité observée peut décroître alors que les mortalités individuelles sont toutes constantes.*

pour un ancien fumeur donné, la probabilité de rechute est constante.

La figure 3 présente une courbe semblable au taux de mortalité observé chez les humains: un fort taux de mortalité infantile, qui diminue jusqu'à la puberté, et qui augmente ensuite. Ce type de courbe ne montre pas forcément que le taux de mortalité individuel diminue puis augmente; on peut l'obtenir même si tous les taux de mortalité individuels augmentent pendant toute la vie. Nous avons obtenu le même résultat en supposant qu'à la naissance, la moitié des bébés sont "faibles" (un fort taux de mortalité, qui augmente avec le temps: $\lambda_1(t) = 0.1 + 0.001t$) et l'autre moitié "forts" (un faible taux de mortalité, qui augmente avec le temps: $\lambda_2(t) = .001 + 0.001t$). Les bébés faibles meurent rapidement, entraînant une prépondérance des bébés forts. Le taux de mortalité chute jusqu'à ce qu'il n'y ait plus que des individus forts, et la courbe suit alors celle de cette sous-population.

La courbe 4 est semblable au taux de divorce: elle augmente jusqu'à un maximum (la "crise des sept ans de mariage"), puis diminue. Comme pour les exemples précédents, il est plausible qu'un modèle homogène rende bien compte de la réalité, et que la plupart des couples suivent effectivement une courbe de cette forme. Mais un autre modèle possible est que dans la plupart des couples, la probabilité instantanée de divorcer augmente de manière affine ($\lambda_1(t) = 0.002 + 0.003t$), mais que quelques couples (ici, 5%) sont particulièrement robustes ($\lambda_2 = 0.002$ constant).

Montrons enfin comment les progrès de la médecine peuvent être mal interprétés. Supposons que la pédiatrie fasse d'énormes progrès, et que le taux de mortalité soit divisé par 2 chez les enfants de moins de 10 ans. Si la population est hétérogène, il est possible qu'on observe la courbe 5 : la mortalité a *augmenté* d'environ 10% pour les personnes de 35 ans. La courbe semble donc montrer

que la médecine a fait des progrès chez les enfants, mais qu'elle a régressé chez les adultes. En réalité, la population était hétérogène ($\lambda_1 = 0.05\exp(0.25t)$; $\lambda_2 = 0.02\exp(0.25t)$; $\pi(0) = 0.7$). Plus d'enfants faibles ont survécu grâce aux progrès de la pédiatrie, et meurent à l'âge adulte.

Bien entendu, nous ne prétendons pas que les explications des courbes que nous proposons soient justes. Des modèles aussi simplistes que les nôtres ne sont certainement pas plus proches de la réalité que les modèles homogènes. Nous avons seulement essayé de souligner que l'approche homogène peut aboutir à des résultats très éloignés de la réalité.

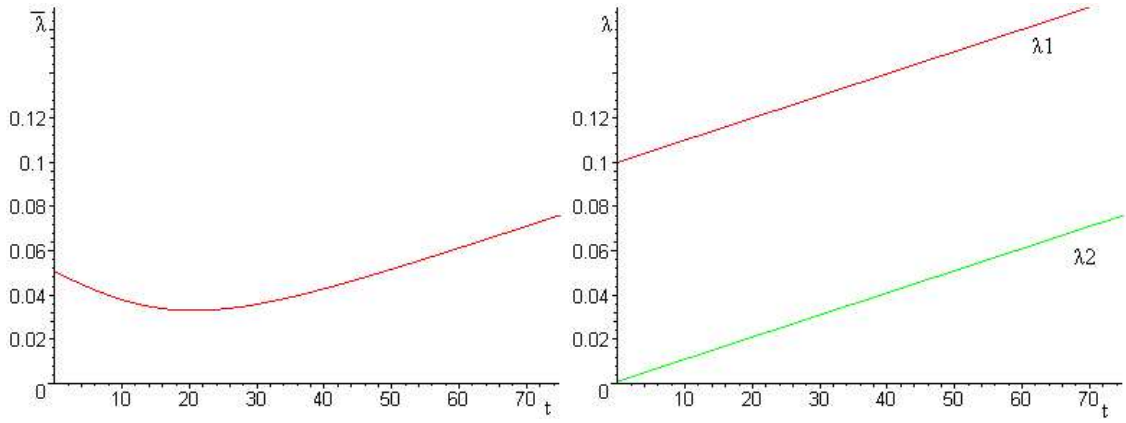

FIG. 3 – *Le taux de mortalité peut baisser avant de remonter même si toutes les mortalités individuelles sont strictement croissantes.*

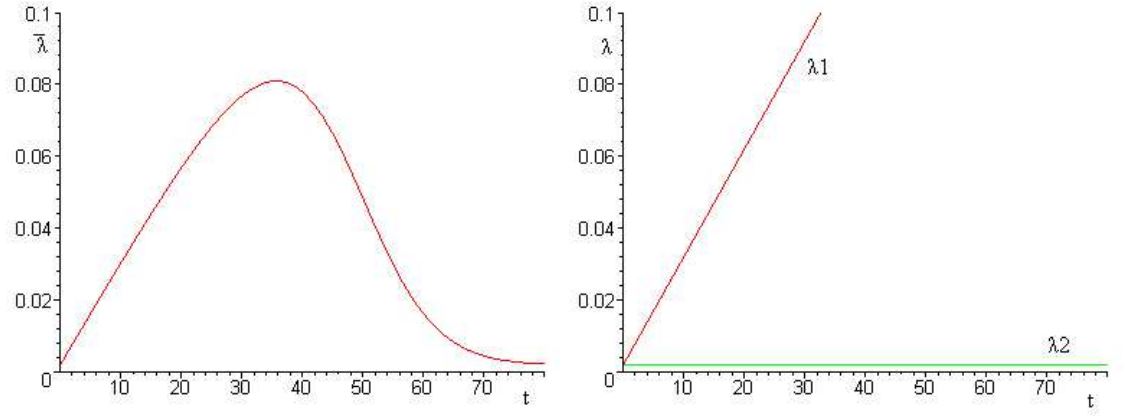

FIG. 4 – *Le taux de mortalité peut augmenter puis diminuer alors que les taux mortalités individuels sont constants ou croissants.*

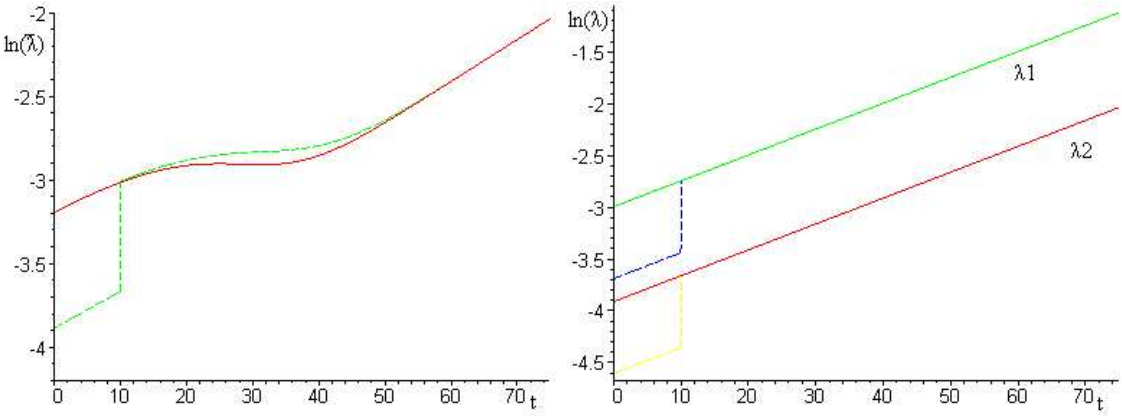

FIG. 5 – *Une diminution de la mortalité chez les enfants peut faire augmenter la mortalité observée chez les adultes.*

Quelle que soit la population considérée, elle sera toujours hétérogène. L'âge, le sexe, l'histoire personnelle, les conditions extérieures auxquelles les individus sont confrontés, etc. varient selon les individus, et peuvent modifier ses chances de mourir. On peut bien sûr essayer de prédire quelles sont les caractéristiques importantes, et espérer obtenir ainsi plusieurs groupes homogènes qu'on étudie séparément. Mais presque toujours, il y aura des sources d'hétérogénéité ignorées. Nous allons donc présenter un modèle qui tient compte de l'hétérogénéité, et qui la modélise par des variables aléatoires.

2 Le modèle de Cox

2.1 Notations

Introduisons tout d'abord de façon plus rigoureuse les outils nécessaires.

Définition 2.1.1 (durée de vie) *La durée de vie d'un individu est une variable aléatoire T positive et continue. On suppose que sa fonction de répartition $F(t) = \mathbb{P}(T \leq t)$ est dérivable.*

C'est cette v.a. qu'on observe.

Définition 2.1.2 (fonction de survie) *La fonction de survie est définie par*

$$S(t) = 1 - F(t) = \mathbb{P}(T > t)$$

Définition 2.1.3 (risque instantané) *La fonction de risque instantané est la fonction*

$$\alpha(t) = \frac{f(t)}{S(t)}$$

où f est la densité de probabilité de T.

Le risque instantané est la probabilité de mourir à t, sachant que l'individu n'est pas mort avant:

$$\alpha(t)\,dt = \frac{f(t)\,dt}{S(t)} = \frac{\mathbb{P}(t < T \leq t + dt)}{\mathbb{P}(T > t)} = \mathbb{P}(t < T \leq t + dt \mid T > t)$$

Définition 2.1.4 (risque cumulé) *La fonction de risque cumulé est donnée par $A(t) = \int_0^t \alpha(u)\,du$.*

La définition de la distribution de T repose que l'une quelconque de ces quatre données, qui sont équivalentes.
On a bien, comme dans la partie 1:

$$S(t) - S(t + dt) = \mathbb{P}(T > t) - \mathbb{P}(T > t + dt) = \mathbb{P}(t < T \leq t + dt) = \alpha(t)S(t)dt$$

et donc $S(t) = \exp(-A(t))$.

2.2 Présentation générale

Le modèle de Cox est le modèle le plus utilisé quand on veut prendre en compte des covariables dans des données de survie.
On considère n individus, p covariables pour chaque individu, et une matrice $\mathbf{X}$ $n \times p$ de covariables. X_{ij} est la covariable j de l'individu i. Le modèle de Cox stipule que pour l'individu i,

$$\lambda_i(t) = \lambda_0(t)e^{\mathbf{X_i}\cdot\beta}$$

où λ_0 est une fonction positive à déterminer, la *fonction de risque de base* commune à tous les individus, et β un vecteur colonne $p \times 1$ de coefficients.
Si on suppose que les covariables ne varient pas avec le temps, le rapport

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{e^{\mathbf{X_i}\cdot\beta}}{e^{\mathbf{X_j}\cdot\beta}}$$

est constant. C'est pourquoi ce modèle est aussi appelé *modèle à risques proportionnels.*
Dans notre cas, il y a trois covariables binaires pour chaque individu, qui indique de quel type de bactérie il s'est nourri (une de trois souches de Esherichia coli considérées: OP50, F11, ED1A). Donc $X_{i1} = 1$ si le ver i s'est nourri de bactéries OP50, et 0 sinon.

L'objectif va être de déterminer la valeur des coefficients qui décrit le mieux la réalité expérimentale. La fonction de risque de base nous intéresse assez peu: ce qu'on veut, c'est pouvoir comparer les différentes covariables. Ici, on veut connaître l'influence du type de bactérie sur le taux de mortalité des vers. Dans la pratique, on va donc fixer un des coefficients à 1 (celui correspondant à ED1A,

pour des raisons purement... alphabétiques), et observer les valeurs des autres coefficients.

Il est à noter que beaucoup de modèles paramétriques traditionnels sont des cas particuliers du modèle (semi-paramétrique) de Cox. Notamment, si $\lambda_0(t) = \lambda$, on obtient le modèle exponentiel, et si $\lambda_0(t) = \alpha\lambda t^{\alpha-1}$, on a le modèle de Weibull.

2.3 Outils mathématiques

On va chercher à estimer les paramètres de notre modèle de façon à ce qu'ils s'approchent au mieux de la réalité. Pour discriminer entre différentes valeurs des paramètres, nous utilisons la notion de *vraisemblance*:

Définition 2.3.1 (vraisemblance) *Soit une famille de probabilités paramétrées, de densité $x \mapsto f(x \mid \beta)$. Sous l'observation x, la vraisemblance est la fonction $L(\beta \mid x) = f(x \mid \beta)$ d'argument β et à x fixé.*

Ainsi, à β fixé, $f(x \mid \beta)$ est la densité de probabilité, et à x fixé, c'est la vraisemblance du paramètre β.

Définition 2.3.2 (estimateur) *Soit $g : \Omega \longrightarrow \mathbb{R}^d$. On appelle estimateur de $g(\beta)$ au vu de l'observation $X = (X_1, \ldots, X_n)$ toute variable aléatoire $\sigma(X)$-mesurable $T : \Omega \longrightarrow \mathbb{R}^d$. Si $\mathbb{E}_\beta(\mid T \mid) < \infty$, on appelle biais de l'estimateur la fonction*

$$\begin{array}{rcl} b_T : \Omega & \longrightarrow & \mathbb{R}^d \\ \beta & \longmapsto & \mathbb{E}_\beta(T) - g(\beta) \end{array}$$

On note $N_i(t) = \mathbb{1}_{\{T_i \leq t\}}$, et $Y_i(t) = \mathbb{1}_{\{T_i \geq t\}}$. Alors:

Proposition 2.3.3 *Le processus stochastique défini par*

$$M_i(t) = N_i(t) - \int_0^t \alpha_i(s) Y_i(s)\, ds \tag{1}$$

est une martingale sous la filtration naturelle $\mathcal{F}_t$.

Démonstration: On a

$$\begin{array}{rcl} \mathbb{E}[dM_i(t) \mid \mathcal{F}_{t-}] & = & \mathbb{E}[dN_i(t) - \alpha_i(t) Y_i(t)\, dt \mid \mathcal{F}_{t-}] \\ & = & \mathbb{P}[dN_i(t) = 1 \mid \mathcal{F}_{t-}] - \alpha_i(t) Y_i(t)\, dt] \\ & = & 0 \end{array}$$

□

On emploie le terme *processus d'intensité* pour désigner $\lambda_i(t) = \alpha_i(t)Y_i(t)$.
On appelle *compensateur* du processus $N_i(t)$ le processus $\Lambda_i(t) = \int_0^y \alpha_i(s)Y_i(s)\,ds$.

Enfin, on définit

$$N_+(t) = \sum_{i=1}^{n} N_i(t)$$

et

$$Y_+(t) = \sum_{i=1}^{n} Y_i(t)$$

qui sont respectivement le nombre total de morts à l'instant t et le nombre total d'individus à risque juste avant l'instant t.

Définition 2.3.4 (estimateur de Nelson-Aalen) *L'estimateur de Nelson-Aalen de la fonction de risque cumulé est défini par*

$$\hat{A}(t) = \int_0^t \frac{\mathbb{1}_{\{Y_+(u)>0\}}}{Y_+(u)}\, dN_+(u)$$

Cet estimateur provient de l'équation (1), qui se réécrit

$$dN_i(t) = \alpha_i(t)Y_i(t)\,dt \;+\; dM_i(t)$$

et où on considère $dM_i(t)$ comme du bruit aléatoire.

Pour étudier cet estimateur asymptotiquement, nous aurons besoin des deux résultats suivants:

Théorème 2.3.5 (Gilvenko-Cantelli) *Pour une suite (X_n) de variables alétaoires, à valeurs dans $\mathbb{R}$, indépendants, de loi F, les répartitions empiriques $\left(\bar{F}_n(\omega,\cdot)\right)$, où*

$$(\bar{F}_n(\omega,\cdot) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(\omega)$$

convergent p.s. uniformément vers F.

Théorème 2.3.6 (Rebolledo) *Si M_n est une suite de martingales telle que*
(i) $\langle M_n\rangle_t$ converge en probabilité vers ν_t déterministe, et
(ii) $\forall\epsilon$, $\exists M_{n,\epsilon}$ suite de martingales telles qu'aucune différence $M_n - M_{n,\epsilon}$ n'ait une amplitude supérieure à ϵ,
alors $M_n(t)$ a une limite $M(t)$ de processus croissant ν_t, et $M(t)$ est un processus gaussien:

$$\frac{M_n(t)}{\nu_y} \xrightarrow{\mathcal{L}} \mathcal{N}(0,\,1).$$

Théorème 2.3.7 (loi asymptotique de l'estimateur de Nelson-Aalen) *La v.a. $\hat{A}(t)$ est un estimateur biaisé de $A(t)$ et, sous l'hypothèse que $A(t) < \infty$, on a*

$$\sqrt{n}\left(\hat{A}(t) - A(t)\right) \xrightarrow{\mathcal{L}} U(t)$$

avec U martingale gaussienne telle que

$$\begin{cases} U(0) = 0 \\ \mathbb{V}(U(t)) = \int_0^t \frac{\alpha(u)}{1-F(u)}\, du \end{cases}$$

Démonstration: Pour le biais, on utilise le fait que $M_+(t)$ est une martingale:

$$\begin{aligned}
\mathbb{E}[\hat{A}(t)] &= \mathbb{E}\left[\int_0^t \frac{\mathbb{1}_{\{Y_+(u)>0\}}}{Y_+(u)}\, dN_+(u)\right] \\
&= \mathbb{E}\left[\int_0^t \frac{\mathbb{1}_{\{Y_+(u)>0\}}}{Y_+(u)}\, \{dM_+(u) + Y_+(u)\alpha(u)\, du\}\right] \\
&= \mathbb{E}\left[\int_0^t \frac{\mathbb{1}_{\{Y_+(u)>0\}}}{Y_+(u)}\, dM_+(u)\right] + \mathbb{E}\left[\int_0^t J(u)\alpha(u)\, du\right] \\
&= 0 + \int_0^t \mathbb{E}[J(u)]\, \alpha(u)\, du \\
&= \int_0^t \mathbb{P}(Y_+(u) > 0)\, \alpha(u)\, du \\
&= \int_0^t \alpha(u)\, du - \int_0^t \mathbb{P}(Y_+(u) = 0)\, \alpha(u)\, du \\
&= A(t) - \int_0^t \mathbb{P}(Y_+(u) = 0)\, \alpha(u)\, du
\end{aligned}$$

Remarquons que ce biais est faible, puisque la probabilité que tous les individus soient morts est proche de zéro.

Pour le caractère asymptotique des résultats, on note

$$Y^n(t) = \sum_{i=1}^n Y_i(t).$$

D'après le théorème de Gilvenko-Cantelli, lorsque $n \longrightarrow \infty$,

$$\sup_{s\in[0,t]} \left| \frac{Y^n}{n} - [1 - F(s)] \right| \xrightarrow{\mathbb{P}} 0$$

Comme

$$\mathbb{1}_{\{Y_+(t)=0\}} = \mathbb{1}_{\{\mathcal{B}(\backslash, \infty - \mathcal{F}(\sqcup))=l\}} \xrightarrow{\mathbb{P}} 0,$$

on a $\mathbb{1}_{\{Y_+(t)>0\}} \xrightarrow{\mathbb{P}} 1$.

Donc

$$\left\langle \sqrt{n}\left[\hat{A}^n - A\right]\right\rangle(t) = \int_0^t n\frac{\mathbb{1}_{\{Y_+(u)>0\}}}{Y^n(u)}\alpha(u)\,du \xrightarrow{\mathbb{P}} \int_0^t \frac{\alpha(s)}{1-F(s)}\,ds$$

qui est déterministe.

Le théorème de Rebolledo donne alors le résultat.

□

2.4 Le modèle de Cox de fragilité

Pour modéliser l'hétérogénéité, on va ajouter une variable aléatoire Θ_i aux covariables dans le modèle de Cox, qui représente l'effet du groupe i. Pour l'individu j du groupe i, le risque instantané s'écrit alors

$$\tilde{\lambda}_{ij} = \lambda_0 \exp(X_{ij}\beta + \Theta_i)$$

Dans cette écriture, X_{ij} est une matrice ligne $1 \times p$. La v.a. Θ_i a généralement pour objectif d'augmenter le risque chez un individu du groupe i. Sa "fragilité" est donc accrue, d'où le nom de *"modèle de fragilité"*. La v.a. Θ_i est la *composante de fragilité* du modèle.
On choisit souvent une loi gaussienne ou une loi gamma pour Θ_i. La variance de cette v.a. est à déterminer, mais on choisit de fixer la valeur moyenne de Θ_i à 1, de façon à ce que $\lambda_0 \exp(X_{ij})$ puisse être identifié au risque d'un individu moyen.

3 Application aux données

3.1 L'algorithme EM

On cherche à calculer les paramètres du modèle de façon à maximiser la vraisemblance $L(\beta \mid \mathbf{X})$ (ou plus précisément la log-vraisemblance $l(\beta \mid \mathbf{X}) = \ln(L(\beta \mid \mathbf{X}))$. Les paramètres que nous cherchons sont les coefficients de β, mais également la variance de la v.a. Θ_i. Dans la suite, nous noterons simplement β le paramètre cherché. L'approche la plus simple est de considérer les réalisations $\theta_i j$ de Θ_i comme des données manquantes.

L'algorithme général de calcul de paramètres avec des données manquantes est *l'algorithme EM*, dont nous donnons un aperçu. La matrice aléatoire des données $\mathbf{Y}$ est composée d'une part de valeurs observables $\mathbf{y_{obs}}$, d'autre part de valeurs manquantes $\mathbf{y_{mqt}}$. La densité d'une v.a. $\mathbf{Z}$ connaissant la valeur (ou la valeur supposée) des paramètres est notée $f(\mathbf{Z} \mid \beta)$, et on note $\ddot{\mathbf{y}}_{\mathbf{obs}}$ le vecteur colonne des durées de vie observées (c'est une réalisation de $\mathbf{y_{obs}}$.
Il faut noter que si β est connu, on connaît la densité de $\mathbf{y_{mqt}}$.
La densité des données complètes se factorise en

$$f(\mathbf{y_{obs}}, \mathbf{y_{mqt}} \mid \beta) = f(\mathbf{y_{obs}} \mid \beta) \times f(\mathbf{y_{mqt}} \mid \mathbf{y_{obs}}, \beta),$$

ce qui montre que

$$f(\mathbf{y_{obs}} \mid \beta) = \frac{f(\mathbf{y_{obs}},\mathbf{y_{mqt}} \mid \beta)}{f(\mathbf{y_{mqt}} \mid \mathbf{y_{obs}}, \beta)}$$

En prenant le logarithme, on obtient la log-vraisemblance pour l'observation $\mathbf{y_{obs}} = \ddot{\mathbf{y}}_{\mathbf{obs}}$:

$$l(\beta \mid \ddot{\mathbf{y}}_{\mathbf{obs}}) = l(\beta \mid \ddot{\mathbf{y}}_{\mathbf{obs}},\mathbf{y_{mqt}}) - \ln(f(\mathbf{y_{mqt}} \mid \mathbf{y_{obs}}, \beta)).$$

On cherche à maximiser cette log-vraisemblance par rapport à β, à $\ddot{\mathbf{y}}_{\mathbf{obs}}$ fixé. Soit $\beta^{\mathbf{t}}$ la t^e approximation de cette valeur recherchée. Comme $l(\beta \mid \ddot{\mathbf{y}}_{\mathbf{obs}})$ est indépendant de $\mathbf{y_{mqt}}$, cette grandeur est égale à son espérance sachant $\mathbf{y_{mqt}}$. Donc pour tout t,

$$l(\beta \mid \ddot{\mathbf{y}}_{\mathbf{obs}}) = \mathbb{E}[l(\beta \mid \ddot{\mathbf{y}}_{\mathbf{obs}})] = Q(\beta \mid \beta^{\mathbf{t}}) - H(\beta \mid \beta^{\mathbf{t}}),$$

où

$$Q(\beta \mid \beta^{\mathbf{t}}) = \int l(\beta \mid \ddot{\mathbf{y}}_{\mathbf{obs}},\mathbf{y_{mqt}}) f(\mathbf{y_{mqt}} \mid \mathbf{y_{obs}}, \beta^{\mathbf{t}})\, d\mathbf{y_{mqt}}$$

et

$$H(\beta \mid \beta^{\mathbf{t}}) = \int \ln(f(\mathbf{y_{mqt}} \mid \mathbf{y_{obs}}, \beta)) f(\mathbf{y_{mqt}} \mid \mathbf{y_{obs}}, \beta^{\mathbf{t}}).$$

L'algorithme EM définit $\beta^{\mathbf{t+1}}$ comme étant la valeur qui maximise $Q(\beta \mid \beta^{\mathbf{t}})$ par rapport à β. On peut montrer, sous des conditions assez générales, que la valeur asymptotique de $\beta^{\mathbf{t}}$ maximise bien $l(\beta \mid \ddot{\mathbf{y}}_{\mathbf{obs}})$. La démonstration étant longue et assez technique, nous nous limitons au résultat suivant:

Théorème 3.1.1 *À chaque itération de l'algorithme EM, la vraisemblance* $l(\beta^{\mathbf{t}} \mid \ddot{\mathbf{y}}_{\mathbf{obs}})$ *est augmentée.*

Démonstration : On cherche à montrer que

$$\forall t,\ l(\beta^{\mathbf{t+1}} \mid \ddot{\mathbf{y}}_{\mathbf{obs}}) \geq l(\beta^{\mathbf{t}} \mid \ddot{\mathbf{y}}_{\mathbf{obs}}).$$

Utilisons pour cela

$$l(\beta \mid \ddot{\mathbf{y}}_{\mathbf{obs}}) = Q(\beta \mid \beta^{\mathbf{t}}) - H(\beta \mid \beta^{\mathbf{t}}).$$

$$l(\beta^{\mathbf{t+1}} \mid \ddot{\mathbf{y}}_{\mathbf{obs}}) - l(\beta^{\mathbf{t}} \mid \ddot{\mathbf{y}}_{\mathbf{obs}}) = [Q(\beta^{\mathbf{t+1}} \mid \beta^{\mathbf{t}}) - Q(\beta^{\mathbf{t}} \mid \beta^{\mathbf{t}})] + [H(\beta^{\mathbf{t}} \mid \beta^{\mathbf{t}}) - H(\beta^{\mathbf{t+1}} \mid \beta^{\mathbf{t}})]$$

Le premier terme entre crochets est positif, car $\beta^{\mathbf{t+1}}$ a été construit pour maximiser $Q(\beta \mid \beta^{\mathbf{t}})$. Montrons que le second terme est également positif.

Notons $f = f(\mathbf{y_{mqt}} \mid \ddot{\mathbf{y}}_{\mathbf{obs}}, \beta)$ et $f^{(t)} = f(\mathbf{y_{mqt}} \mid \ddot{\mathbf{y}}_{\mathbf{obs}}, \beta^t)$.
Soit $Z = Z(\mathbf{y_{mqt}}) = \frac{f}{f^{(t)}}$. Alors, comme

$$H(\beta \mid \beta^{\mathbf{t}}) = \int \ln(f(\mathbf{y_{mqt}} \mid \mathbf{y_{obs}}, \beta)) f(\mathbf{y_{mqt}} \mid \mathbf{y_{obs}}, \beta^{\mathbf{t}}),$$

$$\begin{aligned} H(\beta^{\mathbf{t}} \mid \beta^{\mathbf{t}}) - H(\beta^{\mathbf{t+1}} \mid \beta^{\mathbf{t}}) &= \int \ln\left(\frac{f^{(t)}}{f}\right) f^{(t)}\, d\mathbf{y_{mqt}} \\ &= -\int \ln\left(\frac{f}{f^{(t)}}\right) f^{(t)}\, d\mathbf{y_{mqt}} \\ &= -\mathbb{E}\left[\ln(Z) \mid \ddot{\mathbf{y}}_{\mathbf{obs}}, \beta^{\mathbf{t}}\right] \\ &\geq -\ln\left(\mathbb{E}\left[Z \mid \ddot{\mathbf{y}}_{\mathbf{obs}}, \beta^{\mathbf{t}}\right]\right) = 0 \end{aligned}$$

On a appliqué l'inégalité de Jensen ($-\ln$ est convexe), et utilisé le fait que

$$\mathbb{E}\left[Z \mid \ddot{\mathbf{y}}_{\mathbf{obs}}, \beta^{\mathbf{t}}\right] = \int \frac{f}{f^{(t)}} f^{(t)}\, d\mathbf{y_{mqt}} = \int f\, d\mathbf{y_{mqt}} = 1.$$

□

Le fonctionnement de l'algorithme EM est le suivant:

1. On fixe une valeur initiale $\beta^{\mathbf{0}}$, souvent 0.
2. Étape E (Estimation): On évalue la densité des données manquantes sachant $\beta^{\mathbf{t}}$.
3. Étape M (Maximisation): On calcule $\beta^{(\mathbf{t+1})}$, qui maximise $Q(\beta \mid \beta^{\mathbf{t}})$.
4. On réitère les étapes 2 et 3 jusqu'à obtention de la convergence.

Plus précisément, dans l'étape M, on calcule tous les paramètres sauf la variance η^{t+1}, et dans l'étape E, on calcule η^{t+1} de façon à maximiser la vraisemblance aux autres paramètres fixés à β^{t+1}.
Le terme $Q(\beta \mid \beta^{\mathbf{t}})$ dépend de la fonction de survie, donc des paramètres mais aussi de la fonction de risque de base λ_0, que nous ne connaissons pas, et que nous ne cherchons pas à connaître. On utilise donc l'estimateur de Nelson-Aalen à la place, ce qui est raisonnable si on étudie un grand nombre d'individus.

3.2 Calcul des paramètres et interprétation

Le logiciel statistique R possède une librairie qui permet de calculer automatiquement les paramètres d'un modèle de Cox à l'aide de l'algorithme EM. Nous fournissons donc directement les données proposées par R. Le premier nombre est le coefficient, le second son exponentielle. Rappelons que le coefficient de ED1A a été fixé à 1.
Sans fragilité:

```
>coxph(Surv(Date) Souche, data=donnees)
SoucheF11 1.22 3.382
SoucheOP50 -1.29 0.275
Likelihood ratio test= 272
```

Avec fragilité:

```
>coxph(Surv(Date, Censure) Souche/Rep+frailty(Rep, dist='gauss'), data=donnees)
SoucheF11 1.148 3.152
SoucheOP50 -1.334 0.264
Likelihood ratio test= 300
```

Voici un résultat graphique:

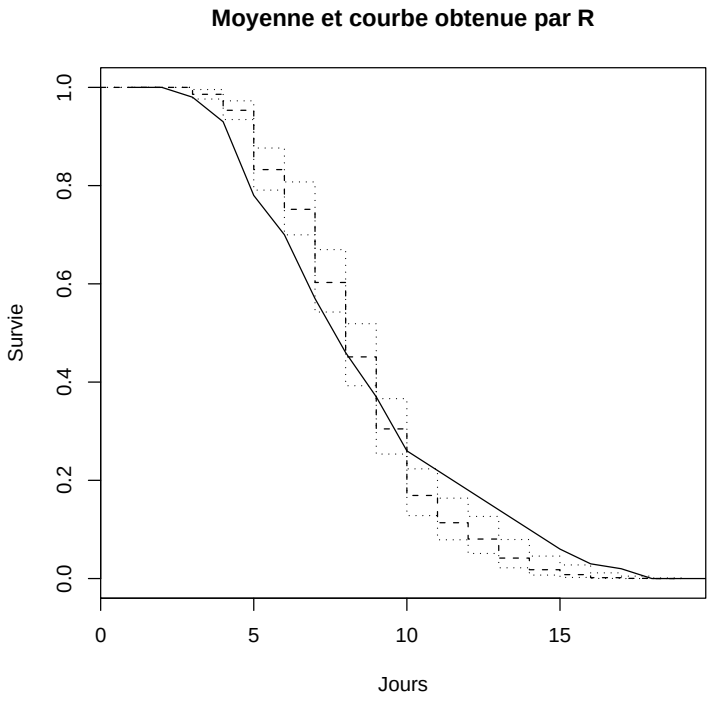

FIG. 6 – *Comparaison des valeurs expérimentales (en trait plein) et du modèle de fragilité de Cox.*

Deux observations principales peuvent être faites:
D'une part, la souche F11 provoque un taux de mortalité 11 fois plus élevé que OP50 (qui est la osuche la plus utilisée en laboratoire), ce qui explique que tous les vers de ces expériences meurent entre le troisième et le dixième jour, alors que c'est entre le quatrième et le vingtième jour pour OP50. De même l'utili-

sation de la souche ED1A multiplie le taux de mortalité par 4 par rapport à OP50.
D'autre part, lorsqu'on omet le terme de fragilité, les coefficients sont surévalués d'environ 7%. Le modèle hétérogène rend mieux compte de la réalité expérimentale, mais le modèle homogène donnait néanmoins une valeur approchée très correcte.

Références

[1] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes.* Springer, 1997.

[2] R. J. Boik. Lecture notes: Statistics 550, Spring 2004. Department of Mathematical Sciences, Montana State University — Bozeman.

[3] S. Brenner. The characteristics of *C. Elegans* that make it useful in teaching. www.loci.wisc.edu/outreach/text/celegans.html.

[4] T. Duchesne. Notes de cours: Analyse des durées de vie, 2004. Université de Laval.

[5] J.-F. Dupuy. *Modélisation conjointe de données longitudinales et de durées de vie.* PhD thesis, Université René Descartes - Paris V, 2002.

[6] T. Lorino. *Modèles statistiques pour des données de survie corrélées.* PhD thesis, Institut National Agronomique Paris-Grignon, 2002.

[7] J. H. Petersen, P. K. Andersen, and R. D. Gill. Variance components models for survival data. *Statistica Neerlandica*, 50:193–211, 1996.

[8] D. Riddle. An introduction to *C. Elegans* for non-specialists. www.biotech.missouri.edu/Dauer-World/Wormintro.html.

[9] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data.* Springer, 2000.

[10] J. W. Vaupel and A. J. Yashin. Heterogeneity's ruses: Some surprising effects of selection on population dynamics. *The American Statistician*, 39(3), August 1985.

A Annexe A: Valeurs expérimentales

Jours	OP50(A)	OP50(B)	OP50(C)	F11(A)	F11(B)	F11(C)	ED1A(A)	ED1A(B)	ED1A(C)
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	3	3	3	0	0	0
4	1	2	1	6	6	7	1	2	3
5	2	4	2	20	21	19	7	9	12
6	3	7	2	31	35	25	8	10	12
7	5	9	7	38	44	34	14	19	19
8	10	11	11	41	48	39	22	29	28
9	13	15	14	42	51	45	32	37	35
10	22	22	18	44	51	45	39	42	44
11	27	27	21	44	51	45	39	45	49
12	30	31	24	44	51	45	39	48	51
13	36	37	29	44	51	45	39	52	51
14	42	42	33	44	51	45	39	52	52
15	45	46	38	44	51	45	39	53	55
16	48	50	44	44	51	45	39	53	55
17	48	53	48	44	51	45	39	53	55
18	48	54	53	44	51	45	39	53	55
19	48	54	55	44	51	45	39	53	55
20	48	54	55	44	51	45	39	53	55
Total	48	54	55	44	51	45	39	53	55

FIG. 7 – *Nombre cumulé de morts pour chacune des expériences*

B Annexe B: Photos de C. Elegans