Modèles de durées

Génia Babykina

evgeniya.babykina@univ-lille.fr

Université de Lille, METRICS
ILIS : Faculté Ingénierie et Management de la Santé

Master BioInfo

Plan

Plan

Plan

Contexte et domaines d'application

Contexte : analyse de durée jusqu'à la survenue d'un événement (analyse de "survie") ou des délais entre les événements ("événements récurrents").

Domaines d'application :

- Médecine : temps jusqu'au décès, temps jusqu'à une rechute, temps entre les ré-hospitalisations, ...
- Biologie : durées entre mise bas des souris, ...
- Economie/sociologie : durées de chômage, les délais entre divorces/re-mariages, les durées entre les conflits militaires,...
- Fiabilité : durée de bon fonctionnement, temps entre les pannes/réparations
- *etc.*

Remarque : événements récurrents (aux instants aléatoires) *vs.* événements répétés (au instants fixés)

Notion d'événement et de "survie"

Survie, mais pas forcément le contexte de "vie et mort"

L'objectif "metier" peut être d'augmenter ou de diminuer le temps jusqu'à un événement → augmenter ou diminuer la *"survie"*

- décès ⇒ augmenter la survie
- échec d'un greffe ⇒ augmenter la "survie"
- guérison ⇒ diminuer la "survie"
- accélérer la recherche de travail ⇒ diminuer le temps jusqu'à l'acquisition ⇒ diminuer la "survie"
- accélérer la prise en charge des patients ⇒ diminuer la "survie"
- *etc.*

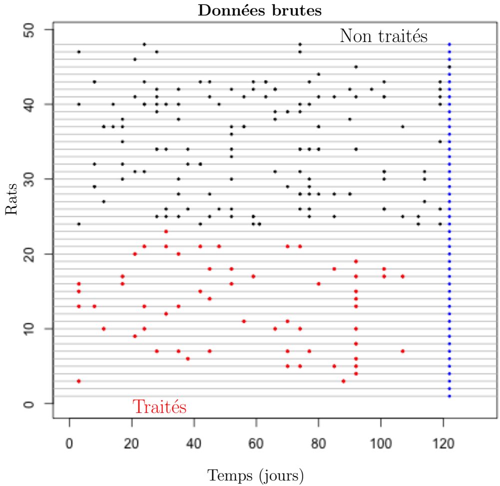
"Survie" : temps jusqu'à l'arrivée d'un événement d'intérêt

Modélisation des durées

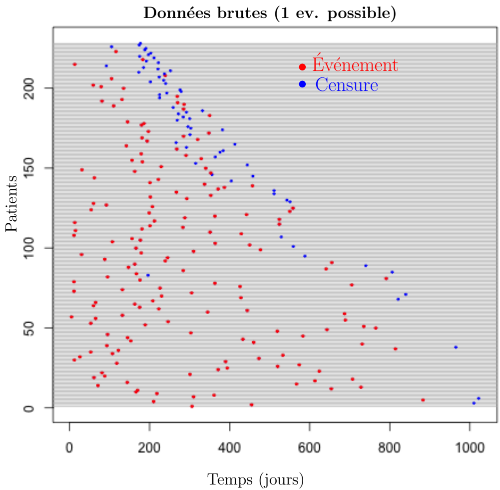
- **Objectifs** :
  - analyser le temps jusqu'à l'occurrence d'un événement
    ⇒ temps jusqu'à la rechute, temps jusqu'à la prise en charge dans les urgences
  - analyser les durées entre les événements
    ⇒ temps entre les épisodes d'épilepsie
  - analyser l'impact de différents facteurs sur la (les) durée(s)
    ⇒ âge, sexe, type de traitement, service de l'hôpital *etc.*
- Pourquoi des **méthodes spécifiques** :
  - comparer le **temps moyen** entre les groupes ⇔ ignorer les perdus de vue, individus sans événement
  - comparer la **fréquence de survenue** d'événements entre les groupes ⇔ ignorer le temps
- **Résultats** de base :
  - Probabilité de "survie" et de "décès" en fonction du temps
  - Survie médiane
  - Impact de différents facteurs sur la "survie"
  - *etc.*

Plan

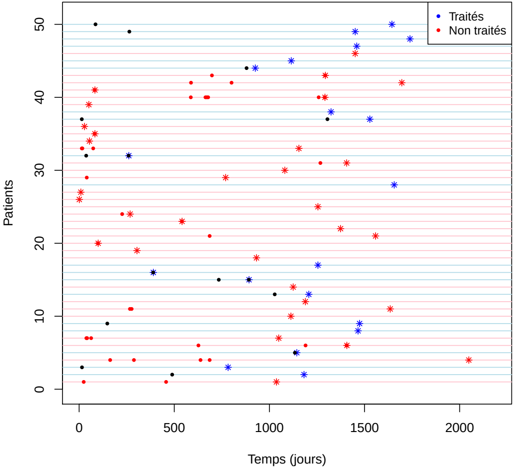
Tumeurs mammaires chez les rats (follow-up 122 jours)



Survie avec cancer des poumons (1 événement : décès)



Ré-hospitalisations des patients après chirurgie



Données sur les re-hospitalisations

Lecture de données

```
load(file="readmission.RData", .GlobalEnv) # Lire les données
head(readmission) # Premières lignes
str(readmission) # Structure de données
```

Tableau de données

```
  id enum t.start t.stop time event      chemo    sex dukes charlson death
1  1    1       0     24   24     1    Treated Female     D        3     0
2  1    2      24    457  433     1    Treated Female     D        0     0
3  1    3     457   1037  580     0    Treated Female     D        0     0
4  2    1       0    489  489     1 NonTreated   Male     C        0     0
5  2    2     489   1182  693     0 NonTreated   Male     C        0     0
6  3    1       0     15   15     1 NonTreated   Male     C        3     0
```

Structure de données

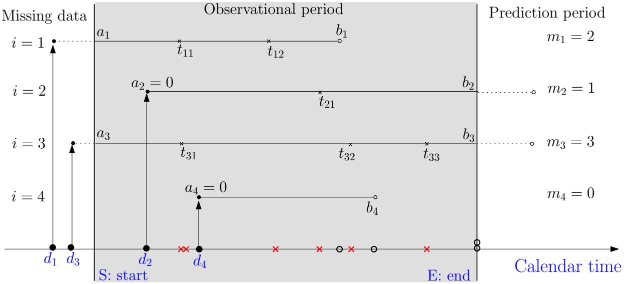
Re-hospitalisations

Patient	Date	Event type	Event number	Age	Sexe	$\cdots$
1	27/01/2001	hospital admission	1	79	F	$\cdots$
1	01/03/2001	end of follow-up	2	79	F	$\cdots$
2	03/02/2000	hospital admission	1	80	H	$\cdots$
2	10/04/2000	hospital admission	2	80	H	$\cdots$
2	23/03/2001	end of follow-up	3	80	H	$\cdots$
...	...	...	...	...	...	$\cdots$

Données : schématiquement

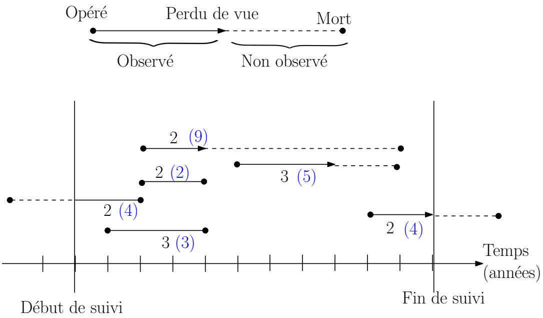
Individu : patient, séjour à l'hôpital, "trajectoire", appareil

Événement : re-hospitalisation, épisode d'asthme, prise en charge aux urgences, panne, *etc.*



Plan

Censure : en tenir compte pour éviter le biais



Durée de vie moyenne réélle $\frac{3+4+4+2+5+9}{6} = 4.5$

Durée de vie moyenne observée $\frac{3+2+2+2+3+2}{6} = 2.33$

- Censure à droite : à la fin d'étude on n'a pas observé l'événement ou à la fin d'étude on n'observe plus l'individu (perdu de vue)
- Autres types de censure : censure à gauche, censure par intervalle

Quelques remarques sur la censure

- Censure par le nombre d'événements :
  - observer tout les individus jusqu'à ce que r entre eux subissent un événement
  - observer tous les individus jusqu'à ce que le nombre d'événement atteinte m
- Censure aléatoire par le temps
  - observer tous les individus jusqu'à un certain instant (fin d'étude)
- Indépendance de censure des temps d'événements est utile mathématiquement ($\Rightarrow$ doit être justifiée)
  - Perte de vue
  - Arrêt de traitement (peut être lié au traitement, censure souvent non indépendante)
  - Fin d'étude : exclus "vivants", censure indépendante

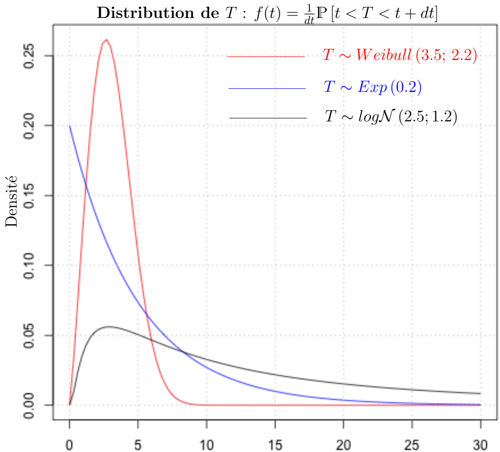
Intuition : indépendance $\Leftrightarrow$ "la probabilité d'avoir un événement est la même pour les individus censurés et non censurés "

Plan

Plan

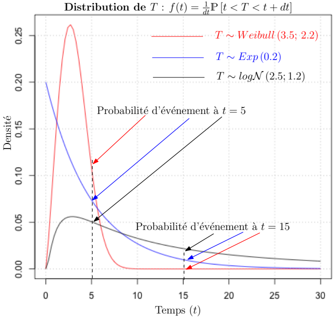
Variable d'intérêt T

Variable : temps jusqu'à un événement : $T \in [0, +\infty[$. Normalité n'est pas adaptée, asymétrie.



Variable d'intérêt T

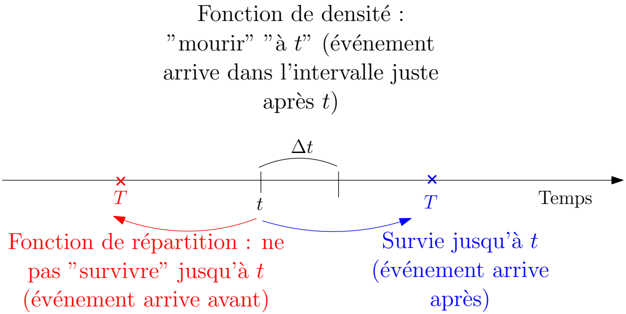
Variable : temps jusqu'à un événement : $T \in [0, +\infty[$. Normalité n'est pas adaptée, asymétrie.



Distributions théoriques avec R

```
curve(dweibull(x,shape=2.2, scale=3.5),
          col="red", from=0, to=30)

curve(dexp(x,rate=0.2), col="blue", add=TRUE)

curve(dlnorm(x, meanlog = 2.5, sdlog = 1.2,
         log = FALSE),
col="black", add=TRUE)
```

Fonctions caractéristiques : illustration



Fonctions caractéristiques : formellement

Survie (événement arrive après t)

$$S(t) = \mathbb{P}\left[T > t\right]$$

Fonction de répartition

$$F(t) = \mathbb{P}\left[T \leq t\right] = 1 - S(t)$$

Fonction de densité

$$\begin{aligned} f(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}\left[t \leq T < t + \Delta t\right]}{\Delta t} \\ &= F'(t) = -S'(t) \end{aligned}$$

$$F(t) = \int_0^t f(u)du$$

Taux d'incidence (risque instantané)

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}\left[t \leq T < t + \Delta t | T > t\right]}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\mathbb{P}\left[t \leq T < t + \Delta t\right]}{\mathbb{P}\left[T > t\right]} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

Hasard cumulé

$$H(t) = \int_0^t h(u)du$$

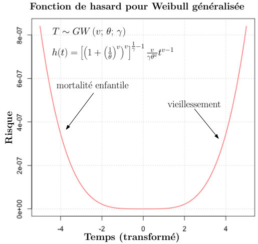
Relation fondamentale

$$S(t) = \exp\left(-H(t)\right) = \exp\left(-\int_0^t h(u)du\right)$$

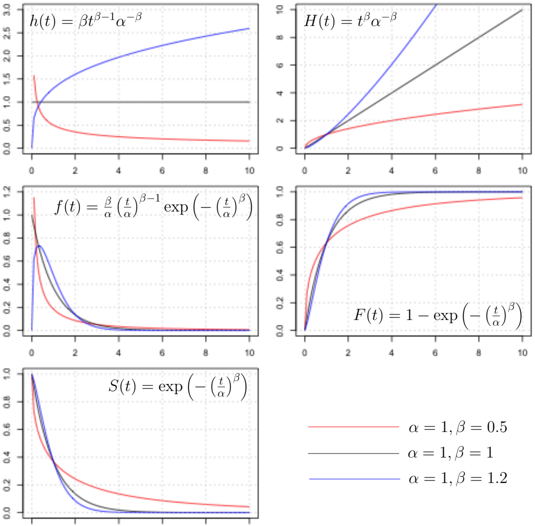
▸ Démonstration

Fonction de hasard $h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq T < t + \Delta t | T > t]}{\Delta t}$

1. Risque constant dans le temps
   - Loi **exponentielle** des durées, $T \sim Exp(\theta)$
   - Processus "sans mémoire" (pas de vieillissement, pas d'accélération d'arrivée d'événements, *etc.*)
2. Risque monotone (diminue ou augmente)
   - Loi de **Weibull** des durées, $T \sim Weibull(v, \theta)$. Loi *Gamma* des durées.
   - Risque augmente, diminue ou reste constant au cours du temps (en fonction de paramètre de forme v)
3. Risque en forme de $\cap$ ou $\cup$ (non monotone)
   - Loi de Weibull généralisée des durées



Exemple $T \sim \mathcal{W}eibull$



Temps de survie moyen et médian

Temps de survie moyen (sans démonstration)

$$\mu = \mathbb{E}(T) = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt \text{ (intégration par partie)}$$

- Pour les durées exponentielles, $f(t) = \lambda \exp(-\lambda t)$, $\mathbb{E}(T) = \frac{1}{\lambda}$, $\text{Med} = \log(2)/\lambda$
- Pour les durées Weibull, $h(t) = \alpha \lambda^\alpha t^{\alpha-1}$, $\mathbb{E}(T) = \frac{\Gamma(1+1/\alpha)}{\lambda}$, $\text{Med} = \frac{\log(2)^{1/\alpha}}{\lambda}$

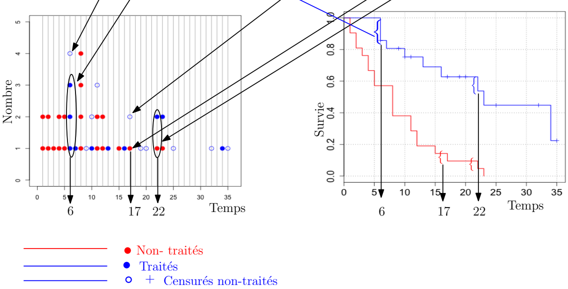
Plan

Construction d'un courbe de survie : illustration

Données : temps jusqu'à l'événement

Patients traités 6∗, 6, 6, 6, 7, 9∗, 10∗, 10, 11∗, 13, 16, 17∗, 19∗, 20∗, 22, 23, 25∗, 32∗, 32∗, 34, 35∗

Patients non-traités 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23



Construction d'un courbe de survie : calculs (1/2)

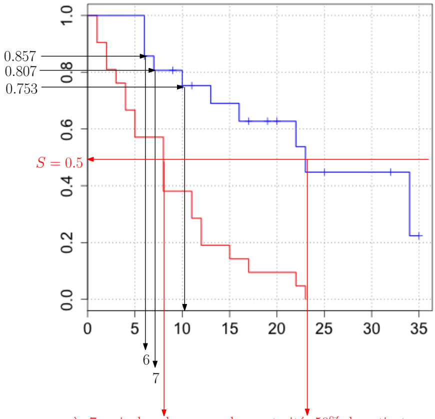
3 décès sur 21 à risque $\Rightarrow S = \underbrace{1 - 3/21}_{} = 0.857$

Temps	Exposés	Censurés	Décès	Survie en t	Survie apres t
6	21	0	3	$\frac{(21-3)}{21}$	0.857
7	17	1	1	$\frac{(17-1)}{17}$	0.807
10	15	1	1	$\frac{(15-1)}{15}$	0.753
13	12	2	1	$\frac{(12-1)}{12}$	0.690
16	11	0	1	$\frac{(11-1)}{11}$	0.627
22	7	3	1	$\frac{(7-1)}{7}$	0.538
23	6	0	1	$\frac{(6-1)}{6}$	0.448
34	2	1	1	$\frac{(2-1)}{2}$	0.224

1 décès sur 17 à risque parmi 85.7% des vivants juste avant $t = 7 \Rightarrow S = 16/17 \times 0.857 = 0.807$

1 décès sur 15 à risque parmi 80.7% des vivants juste avant $t = 10 \Rightarrow S = 14/15 \times 0.807 = 0.753$

Construction d'un courbe de survie : calculs (2/2)



après 7 mois dans le groupe de non traités 50% de patients sont morts, dans le groupe de traités 50% de patients sont morts après 24 mois

Méthode de Kaplan-Meier

Estimation de survie

$$\widehat{S(t)} = \prod_{T_i \leq t} \left(1 - \frac{D_i}{R_i}\right)$$

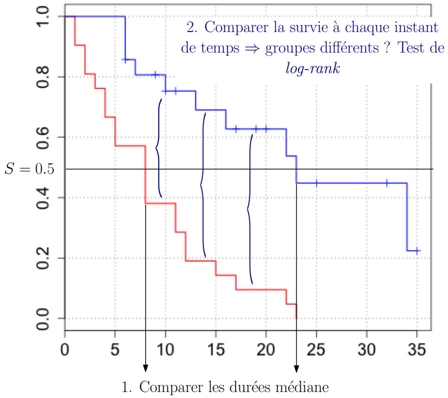
- T_i : instants d'événements, D_i : nombre d'événements à T_i, R_i : nombre à risque à T_i

IC pour survie (Greenwood)

$$\mathrm{se}\widehat{S(t)} = \widehat{S(t)} \sqrt{\sum_{t_i \leq t} \frac{D_i}{(R_i - D_i)\, R_i}}$$

Autre possibilité : méthode actuarielle, Nelson-Aalen estimateur du hasard cumulé : $H(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}, S(t) = \exp(-H(t))$

Comparaison des courbes de survie (1/2)



Comparaison des courbes de survie (2/2)

- Comparaison des durées médianes : si avec le traitement la moitié de patients survivent plus longtemps que sans traitement $\Rightarrow$ le traitement est "efficace"
- Test du log-rank

$$\begin{aligned} \mathrm{H}_0 &: \quad \text{pas de différence entre les courbes}, S_1(t) = S_0(t) \\ \mathrm{H}_1 &: \quad \text{il existe une différence entre deux courbes}, S_1(t) \neq S_0(t) \end{aligned}$$

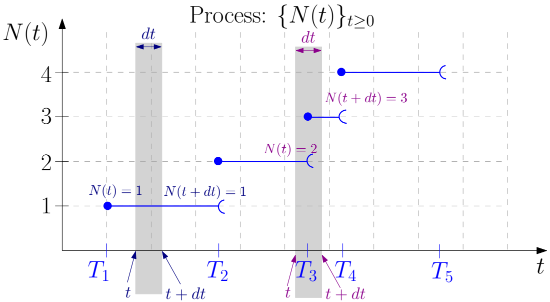
  - principe du test du χ^2 à chaque instant de temps

Test du log-rank

$$u = \frac{\sum_i D_i}{\sqrt{\mathrm{Var}(\sum_i D_i)}} \sim \mathcal{N}(0,1) \text{ sous H0}$$

Plan

Modèle probabiliste : processus de comptage



<u>**Intensité instantanée conditionnelle**</u> : $\lambda(t)$ pour événements récurrents ou fonction de hasard $h(t)$ pour l'analyse de survie

$$\lambda(t) = \lim_{dt\to 0} \frac{1}{dt} \mathbb{P}\left[\underbrace{N(t+dt) - N(t) = 1}_{\text{event in } dt} \,|\, \mathcal{F}(t-)\right]$$

$\mathcal{F}(t)$: historique du processus $\left(N(t), T_1, ..., T_{N(t)}, \boldsymbol{X}(t)\right)$.

Quelques remarques et intuitions

Intensité $\lambda(t)$: fonction réelle positive, intégrable. $\int_0^t \lambda(u)du = \Lambda(t)$.
Processus de comptage $N(t)$ peut être complètement caractérisé par son intensité
Martingale : processus d'espérance 0 et d'incréments indépendants.

L'intensité du processus de comptage par rapport à $\boldsymbol{\mathcal{F}}$

$$\begin{aligned} \lambda(t)\,dt + o(dt) &= \mathbb{P}\left[N(t+dt) - N(t) = 1 \,|\, \mathcal{F}_{t-}\right] \\ &= \mathbb{E}\left[dN(t) \,|\, \mathcal{F}_{t-}\right] \qquad \text{(car } dN(t) \text{ est binaire)} \end{aligned}$$

$o(dt) = \mathbb{P}\left[N(t+dt) - N(t) \geq 2\right]$.

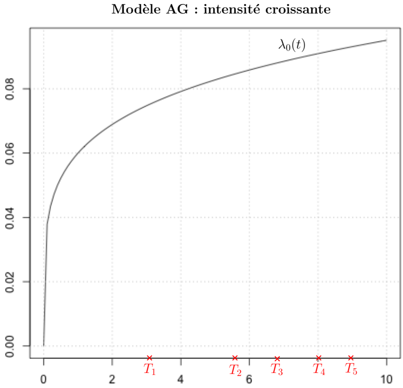
Modèle probabiliste

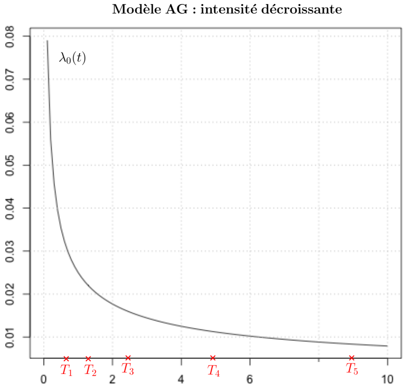
$$\underbrace{N(t)}_{\text{données}} = \underbrace{\int_0^t \lambda(u)\,du}_{\text{modèle}} + \underbrace{M(t)}_{\text{bruit}}$$

$\{\lambda(t)\}_{t\geq 0}$ est l'intensité stochastique de $\boldsymbol{N}$ relativement à $\boldsymbol{\mathcal{F}}$, $\boldsymbol{M} = \{M(t)\}_{t\geq 0}$ est une martingale.

Définir $\lambda(t;\boldsymbol{\theta}) \to$ estimer $\boldsymbol{\theta} \to$ caractériser, prédire $N(t)$

Intuition sur l'intensité



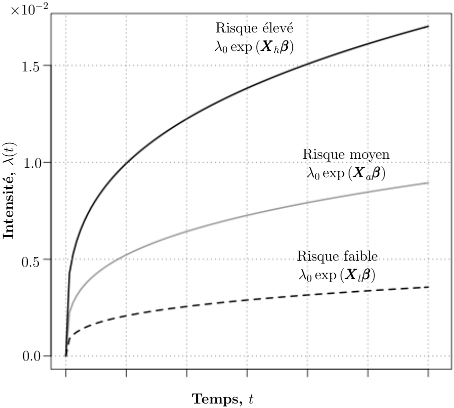


Plan

Modèle (de base) d'intensité (hasard) avec covariables

Modèle de Cox

$$\lambda(t) = \underbrace{\lambda_0(t)}_{\text{Comportement intrinsèque}} \underbrace{\exp\left(X'\beta\right)}_{\text{Stress extérieur, conditions de vie, } etc.}$$



Remarque : temps discret ou continu, $\lambda_0(t)$: spécifiée paramétriquement ou non-paramétriquement

Spécification d'impact de covariables (une des manières !)

Lien linéaire

$$\begin{aligned} \boldsymbol{X}'\boldsymbol{\beta} &= \overbrace{\begin{pmatrix} X_1 & X_2 & \cdots & X_p \end{pmatrix}}^{\text{vecteur de } p \text{ covariables}} \times \overbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_p \end{pmatrix}}^{\substack{\text{coefficients} \\ \text{à estimer}}} \\ &= X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p \end{aligned}$$

Exemple :

$$\lambda(t) = \lambda_0(t)\exp\left(\beta_1 \times Sexe + \beta_2 \times Age + \beta_3 \times Trt\right),$$

avec $Sexe = \begin{cases} 1 & \text{si } homme \\ 0 & \text{si } femme \end{cases}$, $Trt = \begin{cases} 1 & \text{si nouveau} \\ 0 & \text{si existant} \end{cases}$, Age : variable continue

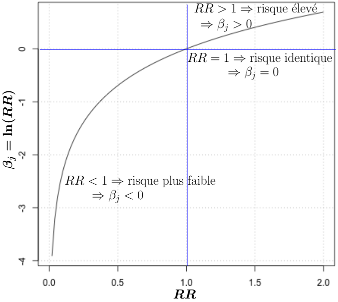
Interprétation d'effet des covariables

Interprétation : effet d'une covariable toute chose égale par ailleurs

$$\lambda(t) = \lambda_0(t) \exp\left(\beta_1 \times Sexe + \beta_2 \times Age\right)$$

Risque relatif RR_h : risque relatif d'événement pour un homme ($Sexe = 1$) *vs.* femme ($Sexe = 0$)

$$\begin{aligned}
RR_h(t) &= \frac{R_h}{R_f} \\
&= \frac{\lambda(t|\text{homme})}{\lambda(t|\text{femme})} \\
&= \frac{\lambda_0(t)\exp(\beta_1 \times 1 + \beta_2 \times (Age = x_2))}{\lambda_0(t)\exp(\beta_1 \times 0 + \beta_2 \times (Age = x_2))} \\
&= \frac{\exp\left(\beta_1 \times 1\right)\exp\left(\beta_2 \times x_2\right)}{\exp\left(\beta_1 \times 0\right)\exp\left(\beta_2 \times x_2\right)} \\
RR_h(t) &= RR_h \quad \text{"proportional hazard" !} \\
&= \exp\left(\beta_1\right) \quad \Rightarrow \beta_1 = \ln RR_h
\end{aligned}$$



Interprétation d'effet des covariables

Interprétation d'un coefficient d'une **variable continue**

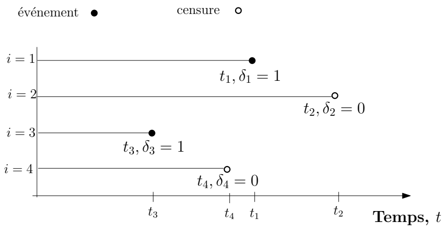
$$\begin{aligned} RR_{age}(t) &= \frac{R_{age1}}{R_{age2}} \\ &= \frac{\lambda(t|\text{age1})}{\lambda(t|\text{age2})} \\ &= \frac{\lambda_0(t)\exp\left(\beta_1 \times (Sexe = x_1) + \beta_2 \times age1\right)}{\lambda_0(t)\exp\left(\beta_1 \times (Sexe = x_1) + \beta_2 \times age2\right)} \\ &= \exp\left(\beta_2\left[age1 - age2\right]\right) \quad \Rightarrow \beta_2\left[age1 - age2\right] = \ln RR_{age} \\ &= R_{age} \end{aligned}$$

e.g. : $5\beta_2$ est le log du risque relatif d'avoir un événement entre une personne de 35 ans et une personne de 30 ans. $\beta_2 > 0 \Rightarrow$ âge est un facteur de risque

Vraisemblance (modèle de Cox)

Vraisemblance : probabilité d'observer ce qu'on observe (données) en fonction de paramètres (à estimer) :

$$\mathcal{L} = f\,(\text{temps d'événements, covariables}, \boldsymbol{\beta}, \boldsymbol{\gamma})$$



on observe : événement à t_1 et à t_3 et censure à t_2 et à t_4

$$\left(\,(t_1, \delta_1)\,,(t_2, \delta_2)\,,(t_3, \delta_3)\,,(t_4, \delta_4)\,\right)$$

$$\mathcal{L} = \mathbb{P}\left[(t_1 \text{ et } \delta_1) \ \text{ et } \ (t_2 \text{ et } \delta_2) \ \text{ et } \ (t_3 \text{ et } \delta_3) \ \text{ et } \ (t_4 \text{ et } \delta_4)\right]$$

Vraisemblance (modèle de Cox)

- **Temps d'événements** $t_i^\star$: réalisation d'une v.a. T de densité $f(t)$ ($\approx$ probabilité de "mourir" à t) et de survie $S(t)$ (probabilité de "survivre" jusqu'à t)
- **Temps de censures** : c_i : réalisation d'une v.a. C de densité $m(t)$ ($\approx$ probabilité d'être censuré à t) et de survie $M(t)$ (probabilité de ne pas être censuré jusqu'à t)
- **On observe** $t_i = \min(c_i, t_i^\star)$, $\delta_i = 1$ si i a un événement, $\delta_i = 0$ si i est censuré

$$
\begin{aligned}
\mathcal{L} &= \mathbb{P}\left[t_1,\delta_1\right]\mathbb{P}\left[t_2,\delta_2\right]\mathbb{P}\left[t_3,\delta_3\right]\mathbb{P}\left[t_4,\delta_4\right] && \text{indépendance des individus}\\
&= f(t_1)M(t_1)\times m(t_2)S(t_2)\times && \text{indépendance entre } T \text{ et } C\\
&\times f(t_3)M(t_3)\times m(t_4)S(t_4) && \text{(censure non-informative)}\\
&= \prod_{i=1}^{n}\left(f(t_i)M(t_i)\right)^{\delta_i}\left(m(t_i)S(t_i)\right)^{1-\delta_i} && \text{pour } n \text{ individus indépendants}\\
&\propto f(t_i)^{\delta_i}S(t_i)^{1-\delta_i} && \text{si censure indépendante !}\\
&= \left(h(t_i)S(t_i)\right)^{\delta_i}S(t_i)^{1-\delta_i} = h(t_i)^{\delta_i}S(t_i)
\end{aligned}
$$

Approche paramétrique

Modèle : pour un individu i

$$h_i(t) = \underbrace{h_0(t)}_{\text{commun}} \underbrace{\exp\left(\boldsymbol{X}_i'\boldsymbol{\beta}\right)}_{\text{individuel}}$$

Idée : risque de base $h_0(t)$ est commun $\Rightarrow$ estimer la forme du risque de base et l'impact des covariables $\boldsymbol{\beta}$

Paramétrique $\Rightarrow$ spécifier la forme (distribution) de $h_0(t)$ et estimer les paramètres de cette distribution

Remarque : estimer $h_0(t) \Leftrightarrow$ estimer $H(t)$, $S(t)$, $f(t)$, obtenir la probabilité d'occurence d'événement en fonction de temps, *etc.*

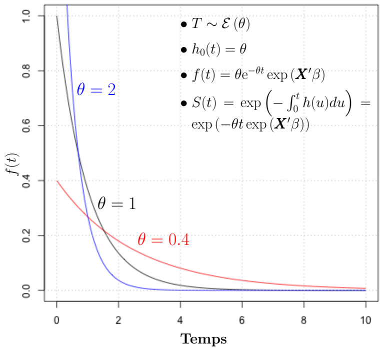
Approche paramétrique : forme de $h_0(t)$

Modèle : pour un individu i

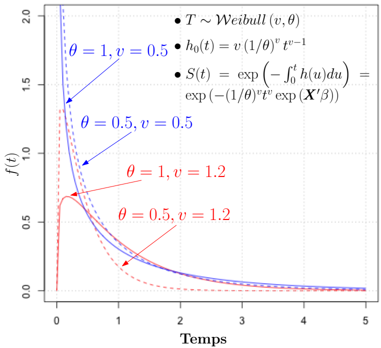
$$h_i(t) = h_0(t)\exp\left(\boldsymbol{X}_i'\boldsymbol{\beta}\right)$$

▸ Détails exponentielle | ▸ Détails Weibull

Loi exponentielle de durée T :



Loi de Weibull de durée T



Modèle de Cox semi-paramétrique

- Modèle $h(t) = h_0(t) \exp(\boldsymbol{X}'\boldsymbol{\beta})$
- Hypothèse : risques proportionnels, $h_0(t)$ commun à tous les individus
- Approche semi-paramétrique : estimation des coefficients $\boldsymbol{\beta}$, h_0 : paramètre de nuisance (non spécifié)
- Estimation (idée) : vraisemblance partielle de Cox :

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{X}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{X}_i'\boldsymbol{\beta})}{\sum_{j \in \mathcal{R}} \exp\left(\boldsymbol{X}_j'\boldsymbol{\beta}\right)} \right)^{\delta_i}$$

  $(\cdots)^{\delta_i}$: sur les individus avec un événement, $\sum_{j \in \mathcal{R}}$: somme sur les individus à risque
- Idée : intervalles sans événement n'ont pas d'information sur $\boldsymbol{\beta} \Rightarrow$ ne tenir compte que des instants d'événements.

Vraisemblance partielle : idée de démonstration

$$\begin{aligned}
\mathbb{P}\left[\text{il y a un ev. à } T_k\right] &= \mathbb{P}\left[(\text{ind. 1 a un ev.}) \textit{ ou } \text{ind. 2 a un ev. } \textit{ou } ...\right] \\
\text{(si indépendance)} &= \mathbb{P}\left[(\text{ind. 1})\ \right] + \mathbb{P}\left[(\text{ind. 2})\ \right] + \cdots \\
\mathcal{R}\left(T_k\right) : \text{à risque à } T_k &= \sum_{j \in \mathcal{R}(T_k)} h_0(T_k) \exp\left(\boldsymbol{X}_j' \boldsymbol{\beta}\right)
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}\left[\text{ind } i \text{ a un ev. à } T_k | \text{il y a un ev. à } T_k\right] &= \frac{\mathbb{P}\left[\text{ind } i \text{ a un ev. à } T_k\right]}{\mathbb{P}\left[\text{il y a un ev. à } T_k\right]} \\
&= \frac{h_0(T_k) \exp\left(\boldsymbol{X}_i' \boldsymbol{\beta}\right)}{\sum_{j \in \mathcal{R}(T_k)} h_0(T_k) \exp\left(\boldsymbol{X}_j' \boldsymbol{\beta}\right)} \\
\text{(si } h_0 \text{ commun)} &= \frac{\exp\left(\boldsymbol{X}_i' \boldsymbol{\beta}\right)}{\sum_{j \in \mathcal{R}(T_k)} \exp\left(\boldsymbol{X}_j' \boldsymbol{\beta}\right)}
\end{aligned}$$

Vraisemblance pour tous les individus $\left(\prod_{i=1}^{n}\right)$ qui ont un événement $()^{\delta_i}$:

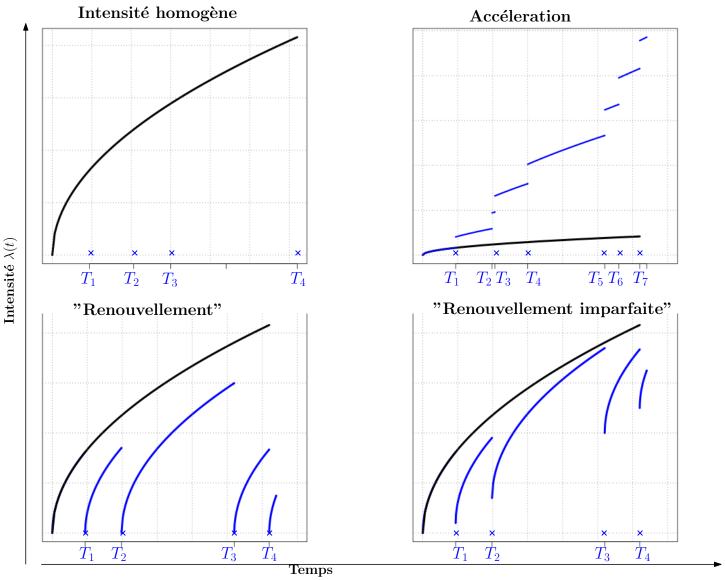
$$\mathcal{L}\left(\boldsymbol{\beta}, \boldsymbol{X}\right) = \prod_{i=1}^{n} \left(\frac{\exp\left(\boldsymbol{X}_i' \boldsymbol{\beta}\right)}{\sum_{j \in \mathcal{R}} \exp\left(\boldsymbol{X}_j' \boldsymbol{\beta}\right)} \right)^{\delta_i}$$

Plan

"Problèmes"

- **Censure** : censure à gauche, censure à droite, par intervalle, indépendance de censure, *etc.*
- **Dépendance entre événements**, accélération/ralentissement de l'intensité : comment modéliser
- **Covariables** : dépendant du temps (stochastiques ou déterministes), grande dimension, *etc.*
- **Problème "théorique"**, propriétés asymptotiques des estimateurs : $n \to \infty$? $m \to \infty$?
- **Problèmes "pratiques"** sur les vraies données
- *etc.*

Dépendance entre les événements



Dépendance entre les événements

Prise en compte par covariables

$$\lambda(t) = \lambda_0(t) \exp\left(\boldsymbol{X\beta} + \gamma N(t-)\right)$$

Modèle à fragilités (⇒ EM pour estimations)

$$\lambda_i(t) = Z_i \lambda_0(t) \exp\left(\boldsymbol{X}_i\boldsymbol{\beta}\right), \quad Z \sim \mathcal{G}(a, b)$$

Âge virtuelle

$$\lambda(t) = \lambda_0(t) \left(A(N(t-), t\right), \quad A(N(t-) : \text{âge à événement } N(t-)$$

Réduction d'intensité

$$\lambda(t) = \lambda_0(t) - f\left(N(t-), T_1, \cdots, T_{N(t-)}; \rho\right)$$

etc. !

En résumé

On cherche à :

- Caractériser le hasard (survie) : quelle forme ? Comment évolue au cours du temps ?
- Durées entre les événements : intensité constante ? intensité d'arrivée d'événements diminue/augmente ?
- Impact des caractéristiques personnelles/environnementales sur la survie/hasard : facteurs de risque ? facteurs protecteurs ?

Plan

Plan

Données de la maternité

Lecture de données

```
load(file="FlowDataAnalyse.RData", .GlobalEnv) # Lire les données
head(data) # Premières lignes
str(data) # Structure de données
```

Tableau de données

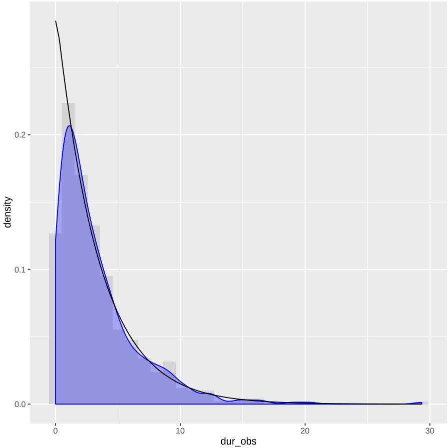
```
    id        Arrival_time           Start_obs             End_obs           Start_lab             End_lab
1 PAT1 2005-04-20 03:22:17 2005-04-20 03:22:17 2005-04-20 03:31:44 2005-04-20 03:31:44 2005-04-20 16:21:40
2 PAT2 2005-04-19 15:12:21 2005-04-19 15:12:21 2005-04-20 03:53:13 2005-04-20 03:53:13 2005-04-21 02:54:18
3 PAT3 2005-04-20 04:17:02 2005-04-20 04:17:02 2005-04-20 06:15:42 2005-04-20 06:15:42 2005-04-20 10:36:31
4 PAT4 2005-04-19 22:13:48 2005-04-19 22:13:48 2005-04-19 22:59:15 2005-04-19 22:59:15 2005-04-20 03:16:28
5 PAT5 2005-04-19 17:47:04 2005-04-19 17:47:04 2005-04-19 21:29:36 2005-04-19 21:29:36 2005-04-20 06:39:53
6 PAT6 2005-04-19 12:01:45 2005-04-19 12:01:45 2005-04-19 17:18:58 2005-04-19 17:18:58 2005-04-20 16:39:41
  Ces_yn            Start_op              End_op            Start_PP              End_PP Cens_yn   Age
1      0                <NA>                <NA> 2005-04-20 16:21:40 2005-04-25 07:20:28       0   <25
2      1 2005-04-21 02:54:18 2005-04-21 05:54:18 2005-04-21 05:54:18 2005-04-26 17:52:09       0   <25
3      0                <NA>                <NA> 2005-04-20 10:36:31 2005-04-21 11:11:21       0 25-35
4      1 2005-04-20 03:16:28 2005-04-20 06:16:28 2005-04-20 06:16:28 2005-04-22 12:32:49       0   >35
5      0                <NA>                <NA> 2005-04-20 06:39:53 2005-04-22 22:06:42       0 25-35
6      1 2005-04-20 16:39:41 2005-04-21 04:44:21 2005-04-21 04:44:21 2005-04-22 06:17:35       0 25-35
```

Calculs sur les données (durées)

```
dur_obs = data$End_obs - data$Start_obs # Durée d'observation pour chaque patiente
dur_obs = as.numeric(dur_obs)*60*60     # Durée en heures
dur_acc = data$End_lab- data$Start_lab  # Durée de travail pour chaque patiente
dur_acc = as.numeric(dur_acc)*60*60     # Durée en heures
etc
```

Analyse de données maternité

Décrire la durée de "séjour" en observation : ajuster une distribution théorique, calculer la durée moyenne, médiane, *etc.*



Données de la maternité

Analyse de "séjour" en observation

```
>hist(dur_obs, probability=TRUE)  # Histogramme
>ff_exp=fitdist(data = dur_obs, distr="exp")  # Ajustement de loi exponentielle
>ff_exp

Fitting of the distribution ' exp ' by maximum likelihood
Parameters:
      estimate Std. Error
rate 0.2985832  0.0133529


>ff_weib=fitdist(data = dur_obs, distr="weibull")  # Ajustement de loi de Weibull
>ff_weib

Fitting of the distribution ' weibull ' by maximum likelihood
Parameters:
      estimate Std. Error
shape 1.005399 0.03480002
scale 3.356201 0.15708034
```

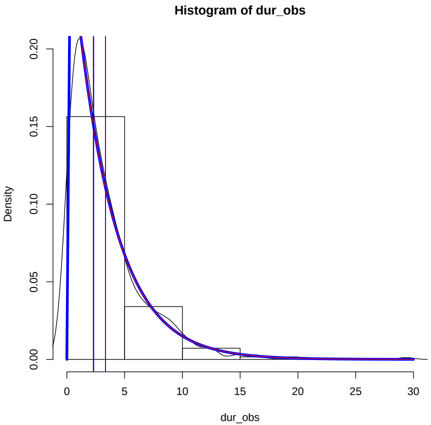
$$\text{Durée} \sim \mathcal{E}xp\,(0.299) \qquad \Rightarrow \mathbb{E}\,[T] = \frac{1}{\text{rate}} = 1/0.299 \approx 3.4\text{heures}$$

$$\text{Durée} \sim \mathcal{W}\,(3.35, 1.00) \qquad \Rightarrow \mathbb{E}\,[T] = \text{scale}\,(1 + \text{shape}) \approx 3.4\text{heures}$$

Données de la maternité

Calcul des moyennes

```
rate=1/ff_exp$estimate
shape=ff_weib$estimate[1]
scale=ff_weib$estimate[2]

# Moyenne observée
mean_obs = mean(dur_obs)
# Espérance de la loi exponentielle
mean_exp=1/rate
# Espérance de Weibull
scale*gamma(1+shape)
# Médiane observée
med_dur=median(dur_obs)
# Médiane de loi exponentielle
med_exp=log(2)/rate
# Médiane de Weibull
med_weib=scale*log(2)^(1/shape)
```



⇒ En moyenne on passe 3.5h dans la salle d'observation, la moitié de patientes en sortent au bout de 2.5h.

Approche semi-paramétrique (modèle de Cox)

Modèle : pour un individu i (modèle de Cox)

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{X}_i'\boldsymbol{\beta}\right)$$

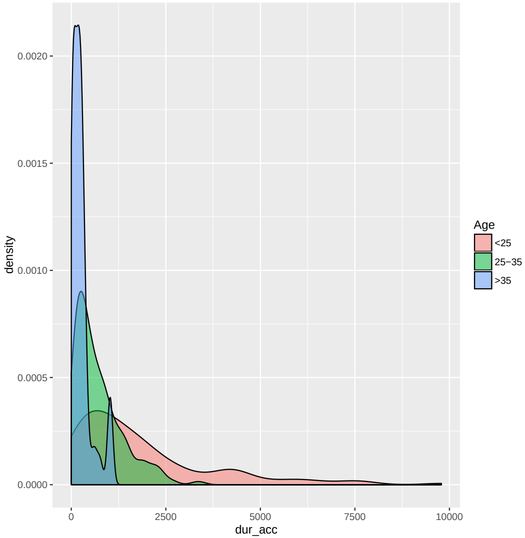
Idée : estimation des coefficients $\boldsymbol{\beta}$, h_0 : paramètre de nuisance (non spécifié) $\Rightarrow$ semi-paramétrique

Analyse des durées d'accouchements : celles qui ont la césarienne au bout d'un certain temps ont des durées **censurées** !

Objet `Surv` : indication de la censure par "+"

```
> Surv(dur_acc, Ces_yn)
 [1]  769.945300+ 1381.082441   260.817113+  257.212377
```

Durées des accouchements *vs.* âge de mère



Cox semi-paramétrique : analyse maternité

$h(t) = h_0(t)\exp(\beta_1 \text{age})$

```
> cox.npar.dur_acc = coxph(Surv(dur_acc, Ces_yn)~Age, data=data)
```

$\forall$ nom — semi-paramétrique — X_1

```
> summary(cox.npar.dur_acc)
Call:
coxph(formula = Surv(dur_acc, Ces_yn) ~ Age, data = data)

  n= 500, number of events= 174
```

	$\hat\beta_j$ coef	RR_{X_j} exp(coef)	$\hat\sigma_{\hat\beta_j}$ se(coef)	Wald ($H_0 : \beta_j = 0$) Z z	p-value Pr(>|z|)	
Age25-35	0.8744	2.3975	0.2090	4.184	2.87e-05	***
Age>35	3.1505	23.3478	0.2551	12.350	< 2e-16	***

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

À tout instant de temps le "risque" est 2.4 fois plus élevé chez les 25-35 ans par rapport aux < 25 ans ⇒ accouchements plus rapides chez les 25-35

À tout instant de temps le "risque" est 23 fois plus élevé chez les > 35 ans par rapport aux < 25 ans ⇒ accouchements beaucoup plus rapides chez les > 35

	exp(coef)	exp(-coef)	lower .95	upper .95
Age25-35	2.397	0.41711	1.592	3.611
Age>35	23.348	0.04283	14.161	38.494

$IC^{95\%}$ pour RR. Si $1 \in IC^{95\%} \Rightarrow$ non significatif

```
Concordance= 0.714  (se = 0.022 )
Rsquare= 0.251   (max possible= 0.975 )
```

qualité globale du modèle

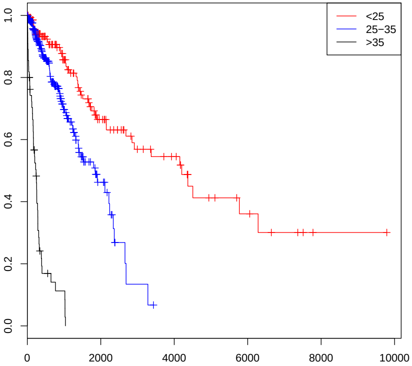
```
Likelihood ratio test= 144.4  on 2 df,   p=0
Wald test            = 178.3  on 2 df,   p=0
Score (logrank) test = 282.7  on 2 df,   p=0
```

tests sur la significativité globale (sur tous les coefficients)

Cox semi-paramétrique : estimation de survie

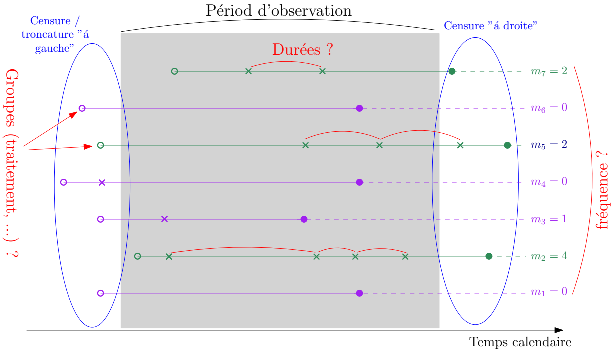
Fonction survfit : estimation de $S(t)$

```
plot(survfit(Surv(dur_acc, Ces_yn)~Age, data=data),
col=c("red", "blue", "black"))
```



Plan

Événements récurrents



Données sur les re-hospitalisations

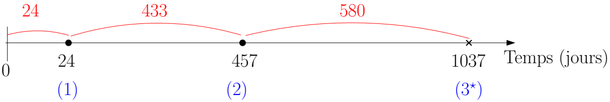
Lecture de données

```
load(file="readmission.RData", .GlobalEnv) # Lire les données
head(readmission) # Premières lignes
str(readmission) # Structure de données
```

Tableau de données

	id	enum	t.start	t.stop	time	event	chemo	sex	dukes	charlson	death
1	1	1	0	24	24	1	Treated	Female	D	3	0
2	1	2	24	457	433	1	Treated	Female	D	0	0
3	1	3	457	1037	580	0	Treated	Female	D	0	0
4	2	1	0	489	489	1	NonTreated	Male	C	0	0
5	2	2	489	1182	693	0	NonTreated	Male	C	0	0
6	3	1	0	15	15	1	NonTreated	Male	C	3	0

Exemple de "trajectoire" d'individu 1 :



Données sur les re-hospitalisations

Variables

```
'data.frame': 861 obs. of  11 variables:
 $ id      : (identifiant) : int  1 1 1 2 2 3 3 4 4 4 ...
 $ enum    : (rank d'événement) int  1 2 3 1 2 1 2 1 2 3 ...
 $ t.start : instant d'événement k : int  0 24 457 0 489 0 15 0 163 288 ...
 $ t.stop  : instant d'événement k+1 : int  24 457 1037 489 1182 15 783 ...
 $ time    : durée entre les événements : int  24 433 580 489 693 15 768 ...
 $ event   : événement ou censure : int  1 1 0 1 0 1 0 1 1 1 ...
 $ chemo   : traitement : Factor w/ 2 levels "NonTreated","Treated": 2 2 2 1 1  ...
 $ sex     : sexe : Factor w/ 2 levels "Male","Female": 2 2 2 1 1 1 1 2 2 2 ...
 $ dukes   : stage de maladie : Factor w/ 3 levels "A-B","C","D": 3 3 3 2 2 2  ...
 $ charlson: indice de comorbidité : Factor w/ 3 levels "0","1-2","3": 3 1 1 1  ...
 $ death   : mort/vivant : int  0 0 0 0 0 0 ...
```

Particularités

Non-indépendance : corrélation entre les événements au sein d'un individu (vs. 1 événement par individu)

Interprétation : hétérogénéité (fragilité) individuelle, dépendance entre les événements (un événement entraîne/retarde les occurrences suivantes)

Conséquences :

- Les effets des covariables sur la durée sont bien estimés : estimateurs des paramètres $\hat{\beta}$ convergents (convergent vers la vraie valeur, non-biaisés) et asymptotiquement normaux ($\Rightarrow$ on peut faire les tests d'hypothèses)
- estimateur de la variance des paramètres $\hat{V} = \hat{\sigma}^2_{\hat{\beta}}$ n'est plus valide $\Rightarrow$ les effets n'apparaissent pas comme "significatifs", on passe à côté des covariables.

Quelques modèles pour les événements récurrents

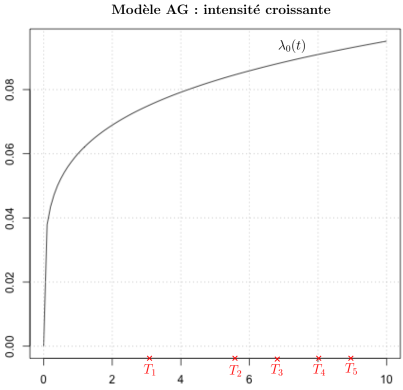
- Modèle d'*Anderson et Gill* (AG)
  - intérêt dans l'estimation de **temps jusqu'à un événement**
  - hypothèse : événements indépendants
  - spécification paramétrique ou semi-paramétrique (comme Cox)
- Modèle de *Prentice-Williams-Peterson gap-time* (PWP-GT)
  - *GT : gap time* : intervalle d'intérêt est la **durée inter-événements**
  - paramétrique ou semi- paramétrique
- Modèle de fragilité (frailty)
  - Idée : certains individus (groupes d'individus) sont plus susceptibles d'avoir les événements que d'autres (plus fragiles) $\Rightarrow$ prendre en compte cette **"fragilité" individuelle**

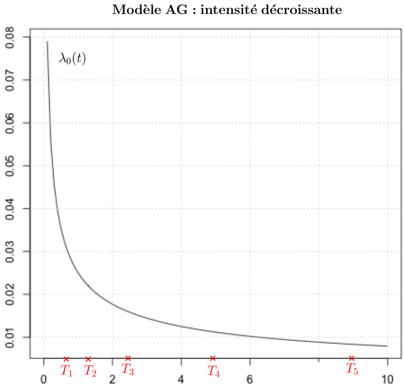
$\rightarrow$ des nombreux autres modèles existent

Modèle AG : intensité d'événements

$$\lambda(t) = \lambda_0(t) \exp\left(\boldsymbol{X}'\boldsymbol{\beta}\right)$$

$\Rightarrow$ Coefficients estimés : β_j (effet de covariable X_j), supposé le même pour chaque événement





AG semi-paramétrique : re-hospitalisations

```
> ag.semipar = coxph(Surv(t.start, t.stop, event) ~ chemo + sex+ dukes+cluster(id), data=readmission)
```

∀ nom

$\lambda(t) = \lambda_0(t)\exp(\beta_1 \text{chemo} + \beta_2 \text{sex} + \beta_3 \text{dukes})$

ev.répétés (corrélation au sein d'id)

```
> summary(ag.semipar)
Call:
coxph(formula = Surv(t.start, t.stop, event) ~ chemo + sex +
    dukes + cluster(id), data = readmission)

  n= 861, number of events= 458
```

T-test ($H_0 : \beta_j = 0$)

	$\hat\beta_{trt}$ coef	RR_{trt} exp(coef)	$\hat\sigma_{\hat\beta_{trt}}$ se(coef)	$\hat\sigma^{recurrent}_{\hat\beta_{trt}}$ robust se	z	p-value Pr(>|z|)	
chemoTreated	-0.2670	0.7657	0.1044	0.1667	-1.602	0.10915	
sexFemale	-0.4994	0.6069	0.1010	0.1685	-2.964	0.00304	**
dukesC	0.3899	1.4768	0.1200	0.1897	2.055	0.03983	*
dukesD	1.5309	4.6225	0.1290	0.2209	6.932	4.16e-12	***

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

	exp(coef)	exp(-coef)	lower .95	upper .95
chemoTreated	0.7657	1.3060	0.5523	1.0615
sexFemale	0.6069	1.6478	0.4362	0.8444
dukesC	1.4768	0.6771	1.0183	2.1418
dukesD	4.6225	0.2163	2.9983	7.1266

$IC^{95\%}$ pour RR. Si $1 \in IC^{95\%} \Rightarrow$ non significatif

```
Concordance= 0.66  (se = 0.014 )
Rsquare= 0.189   (max possible= 0.998 )
```

qualité globale

```
Likelihood ratio test= 180.4  on 4 df,   p=0
Wald test            = 53.76  on 4 df,   p=5.909e-11
Score (logrank) test = 231.1  on 4 df,   p=0,   Robust = 19.76  p=0.0005579
```

tests sur la significativité globale (sur tous les coefficients)

```
  (Note: the likelihood ratio and score tests assume independence of
     observations within a cluster, the Wald and robust score tests do not).
```

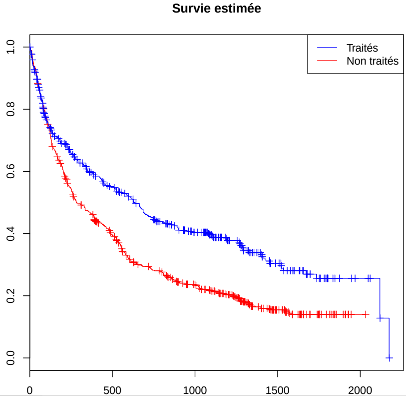
événements répétés ⇒ dépendance !

- traitement diminue le risque de récurrence de 1.3 fois
- risque de récurrence pour les femmes est 1.6 fois moins élevé
- risque de récurrence est 1.4 fois plus élevé pour stade C vs. A-B
- risque de récurrence est 4.6 fois plus élevé pour stade D vs. A-B

AG semi-paramétrique : estimation de survie

Fonction `survfit` : estimation de $S(t)$

```
plot(survfit(Surv(t.start, t.stop, event) ~ chemo+cluster(id)
data=readmission))
```



Modèle conditionnel PWP-GT

Idée : intensité différente pour différents ordres d'événements
⇒ durée entre 2ème et 3ème événements plus courte/plus longue que entre 1er et 2ème

- $k = 1, 2, \ldots$: rang d'événement
- $t - T_{k-1}$: temps écoulé depuis le dernier événement

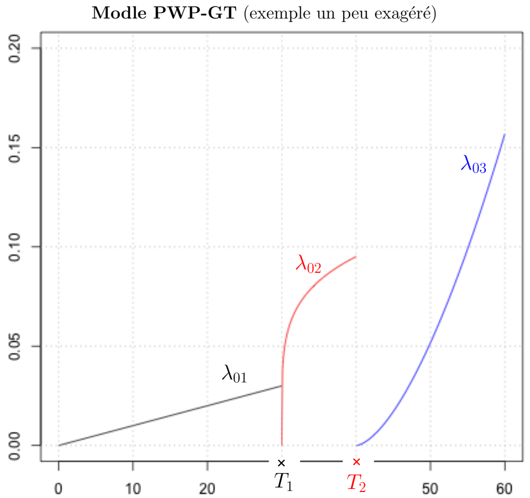
Modèle avec interactions : effet de covariables diffère

$$\lambda_k(t) = \lambda_{0k}(t - T_{k-1}) \exp\left(\boldsymbol{X}'\boldsymbol{\beta}_{\boldsymbol{k}}\right)$$

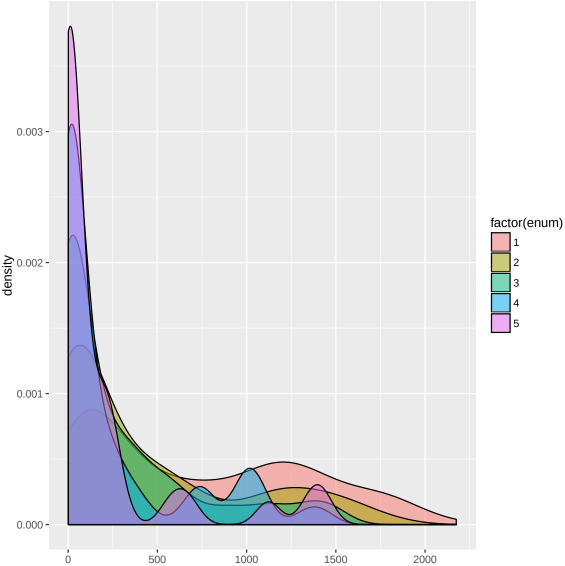
Modèle sans interactions : effet de covariables constant

$$\lambda_k(t) = \lambda_{0k}(t - T_{k-1}) \exp\left(\boldsymbol{X}'\boldsymbol{\beta}\right)$$

Modèle conditionnel PWP-GT



Durées entre re-hospitalisations



PWP-GT sans interactions : re-hospitalisations

$\lambda_{\text{strate}}(t) = \lambda_{0\text{strate}}(t)\exp(\beta_1\text{chemo} + \beta_2\text{sex} + \beta_3\text{dukes})$

```
> pwp = coxph(Surv(t.start, t.stop, event) ~ chemo + sex+ dukes+cluster(id)+
```

stratification par rang d'événement ⇒

```
strata(enum), data=readmission)

> summary(pwp)

Call:
coxph(formula = Surv(t.start, t.stop, event) ~ chemo + sex +
    dukes + cluster(id) + strata(enum), data = readmission)

  n= 861, number of events= 458
```

$\hat\beta_j$ — RR_j — $\hat\sigma_{\hat\beta_j}$ — $\hat\sigma^{\text{recurrent}}_{\hat\beta_j}$ — T-test ($\text{H}_0 : \beta_j = 0$)

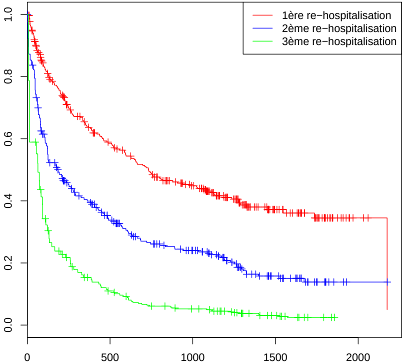
```
                coef exp(coef) se(coef) robust se      z Pr(>|z|)
chemoTreated -0.1985    0.8199   0.1120    0.1185 -1.675  0.09398 .
sexFemale    -0.3384    0.7129   0.1076    0.1113 -3.041  0.00236 **
dukesC        0.2425    1.2744   0.1272    0.1427  1.699  0.08929 .
dukesD        1.0238    2.7839   0.1426    0.1445  7.087 1.37e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

             exp(coef) exp(-coef) lower .95 upper .95
chemoTreated    0.8199     1.2196    0.6499    1.0344
sexFemale       0.7129     1.4027    0.5732    0.8867
dukesC          1.2744     0.7847    0.9635    1.6858
dukesD          2.7839     0.3592    2.0974    3.6951

Concordance= 0.626  (se = 0.024 )
Rsquare= 0.073   (max possible= 0.981 )
Likelihood ratio test= 65.7  on 4 df,   p=1.835e-13
Wald test            = 77.7  on 4 df,   p=5.551e-16
Score (logrank) test = 74  on 4 df,   p=3.22e-15,   Robust = 39.43  p=5.687e-08

  (Note: the likelihood ratio and score tests assume independence of
     observations within a cluster, the Wald and robust score tests do not).
```

- traitement diminue le risque de récurrence de 1.2 fois (vs. 1.3 dans AG)
- risque de récurrence pour les femmes est 1.4 fois moins élevé (vs. 1.6 dans AG)
- risque de récurrence est 1.2 fois plus élevé pour stade C vs. A-B (vs. 1.47 dans AG)
- risque de récurrence est 2.7 fois plus élevé pour stade D vs. A-B (vs. 4.6 dans AG)

PWP sans interactions : estimation de survie

Fonction **survfit** : estimation de $S(t)$

```
pwp = coxph(Surv(t.start, t.stop, event) ~ chemo + sex+ dukes+
    cluster(id)+strata(enum), data=readmission)
plot(survfit(pwp)[1:3], col=c("red", "blue", "green"))
```



PWP-GT avec interactions : R

```
> pwp1 = coxph(Surv(t.start, t.stop, event)~dukes*strata(enum)+cluster(id)+strata(enum),
                                                          data=readmission1)
```

$\lambda_{\text{strate}}(t) = \lambda_{0\text{strate}}(t)\exp(\beta_{1\text{strate}}\text{dukes})$

```
> summary(pwp1)
n= 787, number of events= 405
```

		coef	exp(coef)	se(coef)	robust se	z	Pr(>|z|)	
dukesC	$\hat{\beta}_{\text{C vs. A-B}}$ pour ev.1	0.48430	1.62304	0.16882	0.16912	2.864	0.00419	**
dukesD	$\hat{\beta}_{\text{D vs. A-B}}$ pour ev.1	1.50701	4.51321	0.20970	0.19450	7.748	9.33e-15	***
dukesC:strata(enum)enum=2		-0.21852	0.80371	0.28143	0.29887	-0.731	0.46468	
dukesD:strata(enum)enum=2		-0.80452	0.44730	0.33911	0.34139	-2.357	0.01844	*
dukesC:strata(enum)enum=3		-0.67184	0.51077	0.35518	0.41758	-1.609	0.10764	
dukesD:strata(enum)enum=3		-0.72510	0.48427	0.39964	0.44094	-1.644	0.10009	
dukesC:strata(enum)enum=4		-0.07461	0.92811	0.47436	0.59857	-0.125	0.90081	
dukesD:strata(enum)enum=4		-1.01627	0.36194	0.51210	0.57388	-1.771	0.07658	.
dukesC:strata(enum)enum=5		-1.27335	0.27989	0.67622	0.85772	-1.485	0.13766	
dukesD:strata(enum)enum=5		-1.65137	0.19179	0.79343	0.97484	-1.694	0.09027	.

pas le même effet de covariables $\forall$ strate

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

$\hat{\beta}_{\text{D}}$ pour ev.2 $= 1.50701 - 0.80452 \Rightarrow \exp(0.70248) = 2.018753 = 4.51321 * 0.44730$

$\hat{\beta}_{\text{D}}$ pour ev.4 $= 1.50701 - 1.01627 \Rightarrow \exp(0.49074) = 1.633525 = 4.51321 * 0.36194$

Pour les ré-hospitalisations récurrentes la gravité de tumeur n'a plus grand impact

Plan

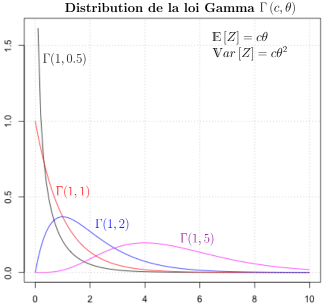
Modèle de fragilité (frailty)

- Idée : certains individus (groupes d'individus) sont plus susceptibles d'avoir les événements que d'autres (plus fragiles)

$\lambda(t) = \lambda_0(t) \exp\left(\boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{W}'\Psi\right)$ $\boldsymbol{W}$ covariables non-observées

$\lambda(t) = \lambda_0(t) \exp\left(\boldsymbol{X}'\boldsymbol{\beta}\right) \exp\left(\boldsymbol{W}'\Psi\right)$

$\lambda(t) = Z\lambda_0(t) \exp\left(\boldsymbol{X}'\boldsymbol{\beta}\right)$ Z : effet aléatoire, souvent $Z \sim \Gamma(1, \theta)$



Modèle de fragilité : R

```
> frailty=coxph(Surv(t.start, t.stop, event)~ chemo + sex + dukes + frailty(id),
                data=readmission)
> summary(frailty)
```

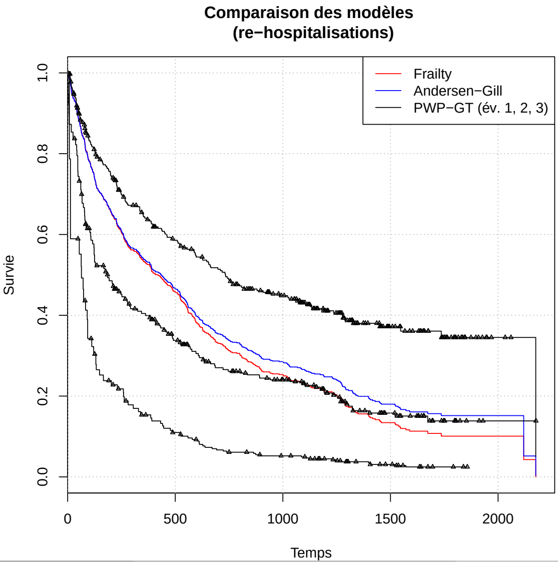
Hétérogénéité individuelle inobservée

```
n= 861, number of events= 458

             coef    se(coef) se2    Chisq  DF    p
chemoTreated -0.2635 0.1704   0.1115   2.39   1.0 1.2e-01
sexFemale    -0.6345 0.1634   0.1081  15.08   1.0 1.0e-04
dukesC        0.3842 0.1907   0.1237   4.06   1.0 4.4e-02
dukesD        1.5437 0.2141   0.1505  51.98   1.0 5.6e-13
frailty(id)                          526.31 223.8 0.0e+00

 Iterations: 5 outer, 28 Newton-Raphson
     Variance of random effect= 1.334991   I-likelihood = -2402.3
 Degrees of freedom for terms=   0.4   0.4   0.9 223.8
 Concordance= 0.854  (se = 0.014 )
 Likelihood ratio test= 839.5  on 225.6 df,   p=0
```

$Var(Z) \neq 0$

```
> ag=coxph(Surv(start, stop, event)~chemo + sex + dukes + cluster
                data=readmission
> summary(ag)

                coef exp(coef) se(coef) robust se      z Pr(>|z|)
chemoTreated -0.2670    0.7657   0.1044    0.1667 -1.602  0.10915
sexFemale    -0.4994    0.6069   0.1010    0.1685 -2.964  0.00304 **
dukesC        0.3899    1.4768   0.1200    0.1897  2.055  0.03983 *
dukesD        1.5309    4.6225   0.1290    0.2209  6.932 4.16e-12 ***
```

Comparaison de modèles



Plan

Analyse des résidus

Résidus martingales : mesure générale de la qualité du modèle, spécification (omission des covariables, *etc.*)

1 événement à t_i : $\hat{r}_i^m = \delta_i(t_i) - \widehat{H}_i(t)$

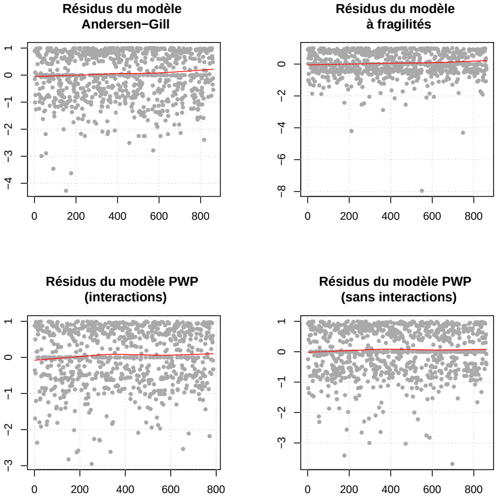
>1 événement à t_{ij} : $\hat{r}_{ij}^m = N_i(t_j) - \widehat{\Lambda}_i(t_j)$

- $\delta_i(t_i)$: événement ou censure (0 ou 1) pour individu i
- $\widehat{H}_i(t)$: hasard cumulé au temps t_i pour individu i
- $N_i(t_j)$: nombre d'événements observé au temps t_j sur l'individu i
- $\widehat{\Lambda}_i(t_j)$: nombre d'événements prédit à t_j pour i.

Intuition :

$$\hat{r}_i^m = 1 - \widehat{H}_i(t) \begin{cases} = 0 & \text{le hasard bien estimé (assez cumulé pour l'événement)} \\ < 0 & \text{"trop de hasard cumulé", événement devait arriver avant} \\ > 0 & \text{"pas assez de hasard cumulé", événement devait pas encore arriver} \end{cases}$$

Analyse des résidus




```
# Résidus du modèle Anderson-Gill
res_ag = resid(ag.semipar,
type='martingale')

# Résidus du modèle de fragilité
res_fr = resid(frailty,
type='martingale')

# Résidus du modèle PWP
#   (avec interactions)
res_pwp_int = resid(pwp1,
type='martingale')

# Résidus du modèle PWP
#  (sans interactions)
res_pwp = resid(pwp,
type='martingale')
```

- doivent être $\approx$ centrés (voir aussi deviance residuals)

Analyse des résidus

Résidus de Schoenfeld : vérification de l'hypothèse des risques proportionnels
$\hat{r}^s_{ik}(t_i)$ calculé pour chaque individu i à l'instant d'événement t_i et pour chaque covariable k

$$\hat{r}^s_{ik}(t_i) = c_i \left(x_{ik} - \hat{x}_{ik}\right)$$

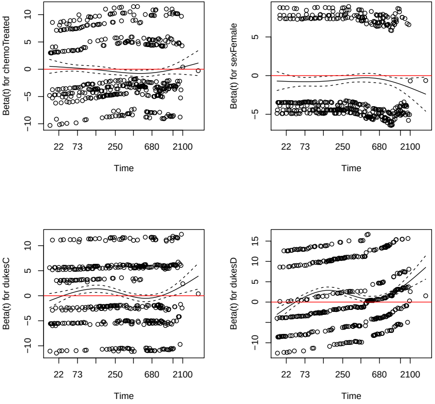
- x_{ik} : valeur observée de covariable
- $\hat{x}_{ik}$: "espérance" de x_{ik}, moyenne de valeur de covariable sur tous les individus à risque à t_i, pondérée par leur probabilité d'avoir l'événement à t_i
- c_i =0 pour les censurés $\Rightarrow$ les résidus ne se calculent que sur les non-censurés

Vérification numérique :

$\mathrm{H}_0 : \beta_j(t) = \beta_j$ effet de covariables est constant dans le temps

$\mathrm{H}_1 : \beta_j(t) \neq \beta_j$ effet de covariables varie dans le temps

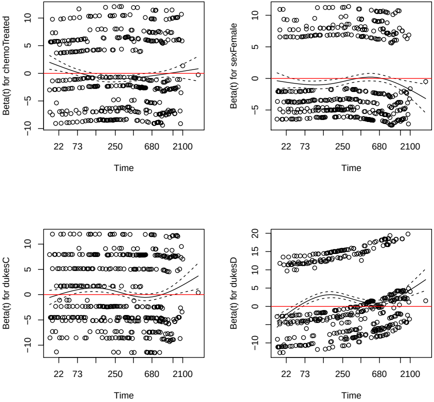
Analyse des résidus : fragilité




```
# Test
> cox.zph(frailty)

                 rho    chisq        p
chemoTreated -0.04029  1.97020 0.160427
sexFemale    -0.00171  0.00349 0.952891
dukesC       -0.00895  0.09532 0.757522
dukesD        0.10014 13.09417 0.000296
GLOBAL             NA 21.03335 0.000312

# Graphique
> plot(cox.zph(frailty))
```

- si hypothèse PH est vérifiée, les résidus de Schoenfeld ont le même pattern selon le temps

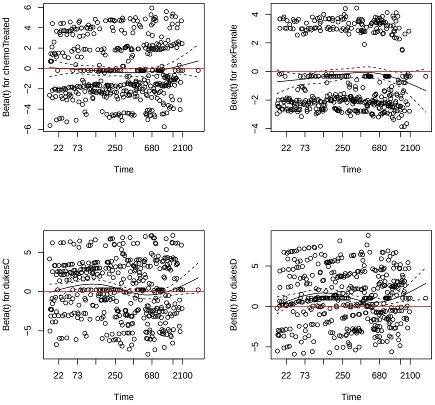
Analyse des résidus : Anderson-Gill




```
# Test
> cox.zph(ag.semipar)


                rho  chisq        p
chemoTreated -0.0453  2.996 0.083495
sexFemale     0.0200  0.522 0.470196
dukesC       -0.0208  0.556 0.455878
dukesD        0.0933 14.738 0.000124
GLOBAL            NA 20.269 0.000442

# Graphique
> plot(cox.zph(ag.semipar))
```

- si hypothèse PH est vérifiée, les résidus de Schoenfeld ont le même pattern selon le temps

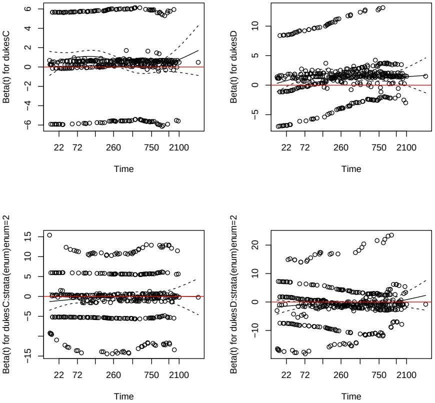
Analyse des résidus : PWP sans interactions




```
# Test
> cox.zph(pwp)



                rho chisq      p
chemoTreated -0.0043 0.010 0.9203
sexFemale     0.0289 0.411 0.5215
dukesC       -0.0723 3.249 0.0715
dukesD       -0.0200 0.196 0.6578
GLOBAL            NA 4.186 0.3815

# Graphique
> plot(cox.zph(pwp))
```

- si hypothèse PH est vérifiée, les résidus de Schoenfeld ont le même pattern selon le temps

Analyse des résidus : PWP avec interactions




```
# Test
> cox.zph(pwp1)

                 rho   chisq      p
dukesC       -0.0810  2.6543 0.1033
dukesD        0.0160  0.0925 0.7610
dukesC:enum=2 0.0581  1.5075 0.2195
dukesD:enum=2 0.0465  0.9459 0.3308
dukesC:enum=3 0.0272  0.4092 0.5224
dukesD:enum=3 0.0264  0.3638 0.5464
dukesC:enum=4 -0.0698 3.9642 0.0465
dukesD:enum=4 -0.0965 4.3039 0.0380
dukesC:enum=5 0.0547  1.9415 0.1635
dukesD:enum=5 0.0374  0.8787 0.3486
GLOBAL            NA 14.8070 0.1393

# Graphique
> plot(cox.zph(pwp1))
```

- si hypothèse PH est vérifiée, les résidus de Schoenfeld ont le même pattern selon le temps

Plan

En R : packages spécifiques

- Ajuster une distribution aux durées observées : package `fitdistrplus`
- Ajuster un modèle de Cox de base : package `survival`
- Modèles paramétriques/à fragilité : package `parfm`
- Modèles à fragilité : package `frailtypack`
- Différents modèles pour les événements récurrents : package `survrec`
- Cox avec time-varying effects : `timereg`
- Visualisation d'importance d'effets des covariables : `rankhazard`
- Modèles de survie paramétriques : `eha, mixPHM`
- Censure par intervalle : `coxinterval`
- et plein d'autres : `http://cran.r-project.org/web/views/Survival.html`

Sujets peu/pas traités

- Autres modèles :
  - de type AG, PWP
  - effet d'intervention sur l'intensité d'événement
  - effet additif des covariables
  - différents modèles de fragilité
  - risques compétitifs
  - *etc.*
- Hypothèses de départ
  - Méthodes de vérification (résidus *etc.*)
  - Modélisation de risque non proportionnel (stratification, covariables dépendant du temps)
  - Modélisation de données avec censure informative
  - *etc.*
- Comparaison de différents modèles
  - critères statistiques AIC, *etc.*
  - "Prédictions" sur les données

Où chercher l'info ?

- Incontournables
  - **Modeling survival data : extending the Cox model**, Therneau, T.M. et Grambsch, P. M., *Springer Science & Business Media*, 2000.
  - 'Survival' package de R
- Packages :
  - Modèles paramétriques/de fragilité : package `parfm`
  - Modèles de fragilité : package `frailtypack`
  - Différents modèles pour les événements récurrents : package `survrec`
  - Cox avec time-varying effects : `timereg`
  - Visualisation d'importance d'effets des covariables : `rankhazard`
  - Modèles de survie paramétriques : `eha, mixPHM`
  - Censure par intervalle : `coxinterval`
  - et plein d'autres (ex. http ://cran.r-project.org/web/views/Survival.html)

Merci !

Questions/comments/suggestions ?

Modèles de durées

Génia Babykina

evgeniya.babykina@univ-lille.fr

Université de Lille, METRICS
ILIS : Faculté Ingénierie et Management de la Santé

Master BioInfo

Lien survie/fonction de hasard $S(t) = \exp\left(-H(t)\right)$

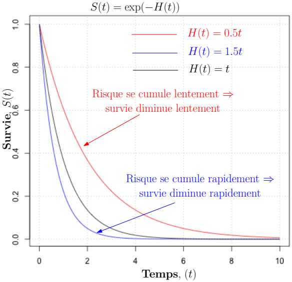
$$f(t) = -S'(t),\ h(t) = \frac{f(t)}{S(t)} \Rightarrow h(t) = -\frac{S'(t)}{S(t)}$$ Retour

Intégration : $\int_0^t \frac{y'(u)}{y(u)} = [\ln y(u)]_0^t \quad , (\ln y(u))' = \frac{1}{y(u)} y'(u)$

$$h(t) = -\frac{S'(t)}{S(t)} \Leftrightarrow \int_0^t \frac{S'(u)}{S(u)} = \int_0^t -h(u) \Leftrightarrow [\ln S(u)]_0^t = -H(t)$$

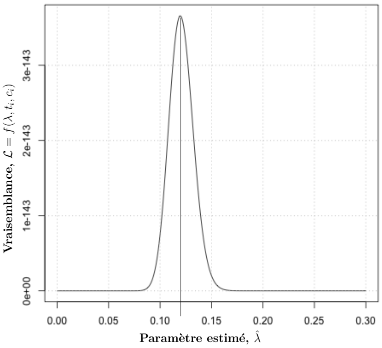
$$\Downarrow$$

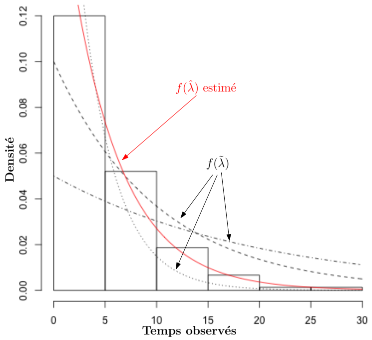
$$\ln S(t) - \ln \underbrace{S(0)}_{=1} = -H(t) \Rightarrow \ln S(t) = -H(t) \Leftarrow S(t) = \exp(-H(t))$$



Fonction de la vraisemblance (loi exponentielle) Retour

- $f(t) = \lambda \exp(-\lambda t)$, $F(t) = 1 - \exp(-\lambda t)$, $S(t) = \exp(-\lambda t)$
- paramètre à estimer : λ
- données : $5.2, 0.9^{\star}, 6.7, 4.5^{\star}, 1.5^{\star}, 0.2, \cdots, n = 150$
- $\propto \prod_{i=1}^{n} (\lambda \exp(-\lambda t_i))^{\delta_i} (\exp(-\lambda t_i))^{1-\delta_i}$





Vraisemblance du modèle exponentiel $h(u) = \theta \exp(\boldsymbol{X}'\boldsymbol{\beta})$, $S(t) = \exp(-\theta t \exp(\boldsymbol{X}'\boldsymbol{\beta}))$

$$
\begin{aligned}
\text{Vraisemblance} = \mathbb{P}[&\text{ev. à } t_1 \text{ pour ind. } 1 \textit{ et } \text{ev. à } t_2 \text{ pour ind. } 2 \textit{ et } \ldots \\
&\textit{et} \text{ pas d'ev. pour ind } 6 \textit{ et } \text{pas d'ev. pour ind. } 9 \textit{ et } \ldots] \\
&= \mathbb{P}\left[\text{ev. à } t_1 \text{ pour ind. } 1\right] \times \mathbb{P}\left[\text{ev. à } t_2 \text{ pour ind. } 2\right] \\
&\times \cdots \\
&\times \mathbb{P}\left[\text{pas d'ev. pour ind. } 6\right] \times \mathbb{P}\left[\text{pas d'ev. pour ind. } 9\right] \\
&\times \mathbb{P}\left[\text{pas d'ev. pour ind. } n\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i) \text{ (cf. slide sur la vraisemblance)} \\
&= \prod_{i=1}^{n} \left(\theta \exp \boldsymbol{X}_i'\boldsymbol{\beta}\right)^{\delta_i} \times \exp\left(-\theta t_i \exp\left(\boldsymbol{X}_i'\boldsymbol{\beta}\right)\right)
\end{aligned}
$$

Retour

Vraisemblance du modèle Weibull $h(t) = v\,(1/\theta)^v\,t^{v-1}\exp(\boldsymbol{X}'\boldsymbol{\beta})$,

$S(t) = \exp\left(-(1/\theta)^v t^v \exp(\boldsymbol{X}'\boldsymbol{\beta})\right)$

$$\begin{aligned}
\text{Vraisemblance} &= \mathbb{P}[\text{ev. à } t_1 \text{ pour ind. 1 } \textit{et} \text{ ev. à } t_2 \text{ pour ind. 2 } \textit{et} \text{ ...} \\
&\quad \textit{et} \text{ pas d'ev. pour ind 6 } \textit{et} \text{ pas d'ev. pour ind. 9 } \textit{et} \text{ ...}] \\
&= \mathbb{P}\left[\text{ev. à } t_1 \text{pour ind. 1}\right] \times \mathbb{P}\left[\text{ev. à } t_2 \text{ pour ind. 2}\right] \\
&\quad \times \cdots \\
&\quad \times \mathbb{P}\left[\text{pas d'ev. pour ind. 6}\right] \times \mathbb{P}\left[\text{pas d'ev. pour ind. 9}\right] \\
&\quad \times \mathbb{P}\left[\text{pas d'ev. pour ind. } n\right]
\end{aligned}$$

$$\mathcal{L} = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i) \text{ (cf. slide sur la vraisemblance)}$$

$$= \prod_{i=1}^{n} \left(v(1/\theta)^v t_i^{v-1} \exp\left(\boldsymbol{X}_i'\boldsymbol{\beta}\right)\right)^{\delta_i} \times \exp\left(-(1/\theta)^v t_i^v \exp\left(\boldsymbol{X}_i'\boldsymbol{\beta}\right)\right)$$

Retour