Analyse de survie : Chapitre 2/2

M. Clertant [1]

Statistiques biomédicales

1. clertant@math.univ-paris13.fr, Lab. Analyse, Géométrie et Applications

Plan

Figure – La hauteur des toboggans représente la proportion estimée des personnes n'ayant pas subi l'événement en fonction du temps.

Comparaison naïve de deux groupes

Deux groupes : **1** et **2**, les estimateurs de Kaplan-Meier des fonctions de survie pour ces deux groupes, $\hat{S}_{\mathbf{1}}(t)$ et $\hat{S}_{\mathbf{2}}(t)$, et leur variance estimées (e.g. Greenwood).

Sous l'hypothèse $H_0 : S_{\mathbf{1}}(t) - S_{\mathbf{2}}(t)$, on a :

$$\frac{\hat{S}_{\mathbf{1}}(t) - \hat{S}_{\mathbf{2}}(t)}{\sqrt{\widehat{Var}(\hat{S}_{\mathbf{1}}(t)) + \widehat{Var}(\hat{S}_{\mathbf{2}}(t))}} \xrightarrow[n_i \to +\infty, i=1,2]{\mathcal{L}} \mathcal{N}(0,1)$$

Cependant, il ne s'agit pas d'une comparaison globale. On ne répond pas à la question :

"Le médicament donné au groupe 1 a-t-il amélioré la survie dans ce groupe par rapport au groupe témoin 2 ?"

On répond à la question :

"La survie a-t-elle était améliorée au temps t ?"

Attention : multiples tests (Bonferroni, une possibilité conservative car sous condition d'indépendance des tests).

Test du log-rank, notations

Pour les groupes **1** et **2**, on considère :

- les temps de décès pour les deux groupes : $T_1 < \ldots < T_n$,

et à chaque temps T_k :

- $d_{\mathbf{1},k}$ et $d_{\mathbf{2},k}$, le nombre de décès dans chacun des groupes ($d_k = d_{\mathbf{1},k} + d_{\mathbf{2},k}$).
- $Y_{\mathbf{1},k}$ et $Y_{\mathbf{2},k}$, le nombre de sujets à risque dans chaque groupes ($Y_k = Y_{\mathbf{1},k} + Y_{\mathbf{2},k}$).

	Décés en T_k	à risque en T_k	Vivant à T_k^+
Groupe **1**	$d_{\mathbf{1},k}$	$Y_{\mathbf{1},k}$	$Y_{\mathbf{1},k} - d_{\mathbf{1},k}$
Groupe **2**	$d_{\mathbf{2},k}$	$Y_{\mathbf{2},k}$	$Y_{\mathbf{2},k} - d_{\mathbf{2},k}$
Ensemble	d_k	Y_k	$Y_k - d_k$

Nombre de sujets à risque en T_{k+1} :

$$Y_k + 1 = Y_k - d_k - c_k,$$

avec c_k, le nombre de censure en T_k (mais celui-ci n'intervient pas dans le calcul, seul le tableau compte).

Loi hypergéométrique

$D_{\mathbf{1},k}$ v.a. hypergéométrique du nombre de morts dans la population, sous la condition $H_0 : S_{\mathbf{1}}(t) = S_{\mathbf{2}}(t)$ et sachant Y_k, $Y_{\mathbf{1},k}$ et d_k :

$$\mathbb{P}(D_{\mathbf{1},k} = d_{\mathbf{1},k}) = \frac{\binom{d_k}{d_{\mathbf{1},k}}\binom{Y_k - d_k}{Y_{\mathbf{1},k} - d_{\mathbf{1},k}}}{\binom{Y_k}{Y_{\mathbf{1},k}}}$$

L'espérance :

$$E_k = \mathbb{E}(D_{\mathbf{1},k}) = \frac{Y_{\mathbf{1},k} \times d_k}{Y_k},$$

et la variance :

$$V_k = Var(D_{\mathbf{1},k}) = \frac{Y_k - d_k}{Y_k - 1} \times \frac{d_k Y_{\mathbf{1},k} Y_{\mathbf{2},k}}{Y_k^2}.$$

Test du log-rank

On compare les réalisations des variables aléatoires $D_{\mathbf{1},k}$ avec leur espérance à chaque temps k. Le test est basé sur la normalité asymptotique de ces différences.

Définition (Test du log-rank)

Sous l'hypothèse $H_0 : S_{\mathbf{1}}(t) = S_{\mathbf{2}}(t)$, on a :

$$\frac{\sum_{k=1}^{n} d_{\mathbf{1},k} - E_k}{\sqrt{\sum_{k=1}^{n} V_k}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

et

$$\frac{\left(\sum_{k=1}^{n} d_{\mathbf{1},k} - E_k\right)^2}{\sum_{k=1}^{n} V_k} \xrightarrow{\mathcal{L}} \chi^2(1),$$

Autres tests

Il est possible d'introduire des poids ω_k dans la statistique :

$$\frac{\sum_{k=1}^{n} \omega_k (d_{\mathbf{1},k} - E_k)}{\sqrt{\sum_{k=1}^{n} \omega_k^2 V_k}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

- **Test de Gehan (Wilcoxon) :** $\omega_k = Y_k$, les décès précoces comptent plus que les décès tardifs.
- **Test de Peto et Prentice :** $\omega_k = \prod_{i=1}^{k} \frac{Y_i}{Y_i + d_k}$, poids plus importants pour les décès précoces.
- **Test du log-rank :** $\omega_k = 1$.

Remarques : 1. on aurait pu définir d'autres statistiques de test équivalentes à l'aide du tableau au slide 5.
2. Les tests sont basés sur l'hypothèse de censure indépendante dans les deux groupes.
3. Perte de puissance du test si les courbes de survie se croisent.

Test du log-rank pour M sous-groupes

Hypothèse nulle $H_0 : S_{\mathbf{1}}(t) = S_{\mathbf{2}}(t) = \dots S_{\mathbf{M}}(t)$.

Pour $1 \leq m \leq M$:

$$E_{\mathbf{m},k} = \mathbb{E}(D_{\mathbf{m},k}) = \frac{Y_{\mathbf{m},k} \times d_k}{Y_k},$$

et la matrice

$$V_{\mathbf{m_1},\mathbf{m_2}} = \frac{Y_k - d_k}{(Y_k - 1)Y_k^2} \times \left[d_k Y_{\mathbf{m_1},k} (Y_{\mathbf{m_2},k} \mathbb{1}_{\{\mathbf{m_1} \neq \mathbf{m_2}\}} + (1 - Y_{\mathbf{m_1},k}) \mathbb{1}_{\{\mathbf{m_1} = \mathbf{m_2}\}}) \right]$$

Soit le vecteur $X^\top = (\sum_{k=1}^{n} d_{\mathbf{1},k} - E_{\mathbf{1},k}, \dots, \sum_{k=1}^{n} d_{\mathbf{m},k} - E_{\mathbf{m},k})$.

On a :

$$X^\top V^{-1} X \xrightarrow{\mathcal{L}} \chi^2(M-1)$$

Remarque : la statistique pour la loi normale est multivariée.

Plan

Distribution de la durée de survie (v.a. continue)

Soit X la v.a. durée de survie ($X > 0$).

La distribution de X est caractérisée par 5 fonctions se déduisant l'une de l'autre :

1. Fonction de répartition (c.d.f.)

$$F(t) = \mathbb{P}(X \leq t), t \geq 0.$$

2. Fonction de survie (survival function)

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

Temps continu : $S(t) = 1 - F(t) = \mathbb{P}(X > t) = \mathbb{P}(X \geq t)$, $(\mathbb{P}(X = t) = 0)$.

Temps discret (heure, jour, mois, année) :
$1 - S^{-}(t) = F^{-}(t) = \mathbb{P}(X < t)$ et $1 - S^{+}(t) = F^{+}(t) = \mathbb{P}(X \leq t)$

Distribution de la durée de survie (v.a. continue)

3. Densité de probabilité f

Soit f, telle que $f(t) \geq 0$ $(\forall t \geq 0)$ et

$$F(t) = \int_0^t f(u)du \quad \text{et} \quad S(t) = \int_t^{+\infty} f(u)du$$

Si F (et S) admet une dérivée :

$$f(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq X \leq t+h)}{h} = F'(t) = -S'(t).$$

$f(t)$ représente la probabilité de mourir dans un petit intervalle de temps après t.

4. Risque instantané λ

La probabilité de mourir dans un petit intervalle de temps après t sachant que l'on a survécu jusqu'au temps t.

$$\lambda(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq X \leq t+h \mid X \geq t)}{h} = \frac{f(t)}{S(t)} = -[\log(S(t)]'$$

5. Taux de hasard cumulé Λ

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log(S(t)).$$

On en déduit que :

$$S(t) = \exp(-\Lambda(t)) = \exp(-\int_0^t \lambda(u)du),$$

et

$$f(t) = \lambda(t)\exp(-\int_0^t \lambda(u)du)$$

Censure à droite indépendante de X

Le cas étudié dans ce cours.

C, la censure aléatoire et X, la durée de survie. On observe :

$$T = X \wedge C \text{ avec } X \perp\!\!\!\perp C \text{ (indépendance),}$$

et aussi δ, la v.a. indiquant si l'événement a été observé ou non (oui : 1, non : 0) :

$$\delta = \mathbb{1}_{\{X \leq C\}}.$$

En temps continue : X et C ont pour densité f et g et pour fonction de survie S et $\overline{G}$.
On a :

$$\mathbb{P}(T \in [t, t+dt], \delta = 1) = \frac{d\mathbb{P}(T \leq t, \delta = 1)}{dt} = f(t)\overline{G}(t)$$

Exercice : 1. Montrer l'égalité ci-dessus.
2. Déterminer une expression équivalente pour $\mathbb{P}(T \in [t, t+dt], \delta = 0)$.

Vraisemblance (Censure dr. ind.)

On observe n individus : $(T_1, \delta_1), \ldots, (T_n, \delta_n)$.

Le modèle f de la durée de survie X dépend d'un paramètre θ.

Dans le cas continue la vraisemblance pour l'individu i est :

$$\begin{aligned}\mathcal{L}_{(T_i,\delta_i)}(\theta) &= \mathbb{P}(T_i \in [t_i, t_i + dt], \delta_i = 1 \mid \theta)^{\delta_i} \times \mathbb{P}(T_i \in [t_i, t_i + dt], \delta_i = 0 \mid \theta)^{1-\delta_i} \\ &= [f(t_i;\theta)\overline{G}(t_i)]^{\delta_i} \times [g(t_i)S(t_i;\theta)]^{1-\delta_i}\end{aligned}$$

Ainsi, la vraisemblance est :

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} \mathcal{L}_{(T_i,\delta_i)}(\theta) = \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i} \times \prod_{i=1}^{n} G(t_i)^{\delta_i} g(t_i)^{1-\delta_i}$$

La partie utile de la vraisemblance est :

$$\mathcal{L}_n(\theta) \propto \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i}.$$

Remarques sur la vraisemblance

Vraisemblance naïve

Sur le sous-échantillon des données non-censurées, la vraisemblance est :

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i}.$$

Le maximum de vrais. a une plus grande variance et est asymptotiquement biaisé.

Censure non-aléatoire (vraisemblance identique)

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i}.$$

Censure à droite de type II

On observe k évts avant arrêt de l'étude :

$$\mathcal{L}_n(\theta) = \binom{n}{k} \times S(t_k;\theta)^{n-k} \times \prod_{i=1}^{k} f(t_i;\theta).$$

Exercice : On suppose que :

- $X \sim \mathcal{E}(\theta)$,
- $C \sim \mathcal{W}(\lambda, \alpha)$.
- La durée de survie est indépendante de la censure.

On observe les données $(t_1, \delta_1), \ldots, (t_n, \delta_n)$. Déterminer le maximum de vraisemblance.

Plan

Modèles à hasards proportionnels

Z vecteur de covariables (par exemple 0 ou 1 pour l'appartenance à un groupe), β, le paramètre d'intérêt.

Le risque instantané prend la forme :

$$\lambda(t \mid Z) = \lambda_0(t) h(\beta^\top Z)$$

Les risques sont proportionnels entre les individus de covariables Z_i et Z_j :

$$\frac{\lambda(t \mid Z_i)}{\lambda(t \mid Z_j)} = \frac{h(\beta^\top Z_i)}{h(\beta^\top Z_j)} \quad \text{(ne dépend pas du temps).}$$

Très souvent, $h(.) = \exp(.)$, car exp positive et $\exp(0) = 1$ ...

Le Hazard Ratio (HR), covariables Z_i et Z_j, $\beta^\star$ vrai paramètre :

$$HR(Z_i, Z_j) = \frac{\lambda(t \mid Z_i)}{\lambda(t \mid Z_j)} = \exp(\beta^{\star\top}(Z_i - Z_j)).$$

Si Z_i et Z_j diffèrent seulement pour cov. k : $HR(Z_i, Z_j) = \exp(\beta_k^\star(Z_{i,k} - Z_{j,k}))$

Si la seule différence est l'appartenance à un groupe (1 pour sujet i, 0 pour sujet j sur cov. k) : $HR(Z_i, Z_j) = \exp(\beta_k^\star)$.

Remarque : $\exp(\beta_k^\star)$ est appelé le hazard ratio pour la coordonnée k. Si

$$Z_i = (Z_1, \ldots, Z_k, \ldots, Z_m) \text{ et } Z_j = (Z_1, \ldots, Z_k + 1, \ldots, Z_m),$$

alors $HR(Z_i, Z_j) = \exp(\beta_k^\star)$

Modèle de Cox

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta^\top Z)$$

- n décès parmi N sujets dans l'étude,
- $1, 2, \ldots, n$ indice des sujets décédés,
- Z_i valeur des covariables de l'individu i,
- $T_1 < T_2 < \ldots < T_n$ temps des décès observés,
- R_i, individu à risque en T_i^-.

Question : Exprimer $S(t|Z)$ dans le modèle de Cox.

Vraisemblance partielle

Intuition :

- on s'intéresse uniquement à β (paramètre d'intérêt),
- on ne cherche pas à estimer λ_0 (paramètre de nuisance),
- aucune information donnée sur β pendant les intervalles où l'on observe pas de décès (c'est une approximation : la censure pourrait en apporter, mais peu),
- on travaille donc conditionnellement au temps de décès T_i.

Sachant l'ens. de sujets à risque R_i, la probabilité de décès dans $[T_i, T_i + dt]$ est :

$$\sum_{j \in R_i} \lambda_0(T_i) \exp(\beta^\top Z_j).$$

La probabilité que ce soit l'individu i :

$$\frac{\lambda_0(T_i) \exp(\beta^\top Z_i)}{\sum_{j \in R_i} \lambda_0(T_i) \exp(\beta^\top Z_j)} = \frac{\exp(\beta^\top Z_i)}{\sum_{j \in R_i} \exp(\beta^\top Z_j)}.$$

Et finalement la vraisemblance partielle de Cox :

$$L_{cox}(\beta) = \prod_{i=1}^{n} \frac{\exp(\beta^\top Z_i)}{\sum_{j \in R_i} \exp(\beta^\top Z_j)}.$$

Log-vraisemblance partielle :

$$\mathcal{L}(\beta) = \log(L_{cox}(\beta))$$

Question : Déterminer le vecteur $U(\beta)$ des dérivées partielles de la log-vraisemblance.

L'estimateur de Cox, $\hat{\beta}$, satisfait :

$$U(\beta) = 0.$$

Pas de solution exacte à cette équation ; approximation possible avec l'algo. de Newton-Raphson

La matrice d'information associée au modèle de Cox est $I_n(\beta)$:

$$[I_n(\beta)]_{i,j} = -\frac{\partial^2 \mathcal{L}(\beta)}{\partial\beta_i \partial\beta_j}.$$

Un estimateur de $Var(\hat{\beta})$ est :

$$\widehat{Var}(\hat{\beta}) = I_n(\hat{\beta})^{-1}$$

De façon très informelle, comment pourrait-on justifier cet estimateur ?

Rappel

Soient $X_1, \ldots, X_n$ des observations iid suivant la loi f_θ : $X_k \sim f_\theta$.
L'estimateur du maximum de vraisemblance conditionnellement aux X_i est $\hat{\theta}_n$.

Normalité asymptotique de l'emv

Sous des conditions de régularité, on a :

$$\sqrt{n}(\hat{\theta}_n - \theta) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, I(\theta)^{-1}),$$

où $I(\theta)$ est l'information de Fisher : $I(\theta) = -\mathbb{E}\left[\dfrac{\partial^2}{\partial\theta^2}\log(f_\theta(X))\right]$.

On pose : $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$ et $\theta_0 = (\theta_1, \ldots, \theta_{k'}) \in \Theta_0$ avec $k' < k$ de sorte que f_{θ_0} est un sous-modèle de f_θ : $\Theta_0 \subset \Theta$. On note L la fonction de vraisemblance et $\hat{\theta}$ et $\hat{\theta}_0$ les emv respectifs.

Théorème de Wilks

Sous des conditions de régularité et θ étant le vrai paramètre, si l'hypothèse $H_0 : \theta \in \Theta_0$ est vraie, on a :

$$-2\log\left(\frac{L(\hat{\theta}_0)}{L(\hat{\theta})}\right) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \chi^2(k-k').$$

Question : On étudie le taux de rechute, pour deux groupes, le premier ayant bénéficié d'une opération préventive ($Z = 0$), le second réunissant les patients qui n'ont pas eu d'opération mais bénéficie du même traitement ($Z = 1$). Le risque instantané est estimé par un modèle de Cox.

Un premier essai clinique à large échelle (n=10000) réunissant des patients adultes a permis d'ajuster le paramètre β : $\beta^\star \simeq 0.04$.

Un second essai clinique se fait sur une population de patients de plus de 60 ans (n=100). La fonction de vraisemblance est $\mathcal{L}_{cox,2}$. Le paramètre β est estimé par : $\hat{\beta}_{cox,2} \simeq 0.5$. L'estimateur de la variance est $I_{n,2}(0.5) = 30$. La fonction de vraisemblance appliquée aux deux valeurs estimées de β donne : $\mathcal{L}_{cox,2}(0.5) = 0.04$ et $\mathcal{L}_{cox,2}(0.04) = 0.0002$.

Proposer deux méthodes pour tester l'hypothèse H_0, la première estimation de $\beta^\star = 0.04$ était la bonne pour les patients de plus de 60 ans.

Distribution asymptotique

$\beta^\star$, le vrai paramètre.

Distributions pour $\hat{\beta}$

Quand n tend vers l'infini :

$$I_n(\hat{\beta})(\hat{\beta} - \beta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

et

$$(\hat{\beta} - \beta^\star)^\top I_n(\hat{\beta})(\hat{\beta} - \beta^\star) \xrightarrow{\mathcal{L}} \chi^2(p).$$

Distributions pour le rapport des vraisemblances partielles

$$2\left(\log \mathcal{L}_{cox}(\hat{\beta}) - \log \mathcal{L}_{cox}(\beta^\star)\right) \xrightarrow{\mathcal{L}} \chi^2(p).$$

Tests

Hypothèse nulle : $H_0 : \beta = \beta_0$ et β vecteur de dimension p.

Statistique du rapport des vraisemblances

$$2\left(\log \mathcal{L}_{cox}(\hat{\beta}) - \log \mathcal{L}_{cox}(\beta_0)\right) \xrightarrow{\mathcal{L}} \chi^2(p).$$

Statistique de Wald

$$(\hat{\beta} - \beta_0)^\top I(\hat{\beta})(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \chi^2(p).$$

Statistique du score

$$(U(\beta_0))^\top I(\beta_0)^{-1}(U(\beta_0)) \xrightarrow{\mathcal{L}} \chi^2(p).$$

Test univarié

On teste un modèle avec une covariable de plus Z_{p+1}. Celle-ci est-elle utile ?
L'hypothèse nulle est : $H_0 : \beta = \beta_0$, avec

$$\beta_0 = (\beta_1, \ldots, \beta_p, 0).$$

La statistique de Wald est :

$$(\hat{\beta} - \hat{\beta}_0)^\top I(\hat{\beta})^{-1}(\hat{\beta} - \hat{\beta}_0) \xrightarrow{\mathcal{L}} \chi^2(1)$$

avec $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\beta}_{p+1})$ et $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p, 0)$

Cela revient à rejeter H_0 quand la statistisque

$$\hat{\beta}_{p+1} / I^2_{j,j}(\hat{\beta})$$

est plus grande que le quantile d'ordre $1 - \alpha$ d'une loi du χ^2 à un degré de liberté.

Estimation du risque cumulé de base λ_0

L'estimateur de Breslow est :

$$\hat{\Lambda}_0(t) = \sum_{i:T_i \leq t} \frac{d_i}{\sum_{j \in R_i} \exp(\hat{\beta}^\top Z_j)},$$

avec d_i le nombre de décès en T_i, c'est un extension de l'estimateur de Nelson-Aalen. Si $\hat{\beta} = 0$, on retrouve l'estimateur de Nelson-Aalen.

Question : En déduire un estimateur de la fonction de survie.

Adéquation du modèle (1/2)

Pour deux covariables Z_i et Z_j, dans le modèle de Cox le risque relatif est :

$$RR(Z_i, Z_j) = \frac{\lambda(t \mid Z_i)}{\lambda(T \mid Z_j)} = \exp[\beta^\top (Z_i - Z_j)].$$

Le risque relatif est constant au cours du temps est log-linéaire.

Méthode graphique

Covariable à deux modalités (1 et 2) ; estimation de la fonction de survie dans chacun des groupes ($\hat{S}_1$ et$\hat{S}_2$).
Les courbes des fonction $\log[-log(\hat{S}_i(t))]$ doivent être parallèles, car dans le modèle de Cox :

$$\log[-log(\hat{S}_i(t))] = \log(\hat{\Lambda}_0(t)) + \hat{\beta} Z.$$

Variation au cours du temps

Interaction entre le temps et la covariable :

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta Z + \gamma Z \log(t))$$

Pour une covariable binaire : $RR(1, 0) = e^\beta \times t^\gamma$.
Idée : Tester la nullité de γ.

Adéquation du modèle (2/2)

Résidus de Cox-Snell

On a :

$$S(x) = \exp(-\Lambda(x \mid Z)) \text{ avec } \Lambda(x \mid Z) = \Lambda_0(x)\exp(\beta^\top Z)$$

La variable $Y = \Lambda(X \mid Z)$ vérifie :

$$\mathbb{P}(Y > y) = \mathbb{P}(X > \Lambda^{-1}(y \mid Z)) = \exp(-y)$$

Adéquation globale du modèle si le risque cumulé est proche de la droite y=x (risque cumulé pour $\mathcal{E}(1)$.).
En pratique :

- On estime $\Lambda(. \mid Z)$ par $\hat{\Lambda}_0(.)\exp(\hat{\beta}^\top Z)$.
- On introduit les variables $v_i = \hat{\Lambda}_0(T_i)\exp(\hat{\beta}^\top Z)$, $(i = 1, \ldots, n)$.
- On estime le risque cumulé des v_i par l'estimateur de Nelson-Aalen $\hat{\Lambda}_v$.
- La courbe $y = \hat{\Lambda}_v(x)$ doit être proche de la droite d'équation $y = x$.

Plan

Modèle de Cox stratifié

Il est possible de stratifier le modèle en fonction de groupes. Par exemple, les enfants et les adultes (Y=0,1). Les covariables jouent alors le même rôle dans chacun des groupes mais la fonction de risque de base est différente.
Pour les enfants :

$$\lambda(t \mid Z, Y = 0) = \lambda_0(t) \exp(\beta^\top Z).$$

Pour les adultes :

$$\lambda(t \mid Z, Y = 1) = \lambda_1(t) \exp(\beta^\top Z).$$

Hypothèse : Même utilisation des covariables dans chaque strates.

Test de l'hypothèse (rapport de vrais.)

$$2 \left[\log(L_{Cox,0}(\hat{\beta}_0) \times L_{Cox,1}(\hat{\beta}_1)) - \log(L_{Cox}(\hat{\beta}))\right] \xrightarrow{\mathcal{L}} \chi^2(p),$$

où $L_{Cox,i}$ et $\hat{\beta}_i$ $(i = 0, 1)$ sont les vraisemblances et leurs estimateurs pour les strates 0 et 1, et L_{Cox} et $\hat{\beta}$ la vraisemblance et son estimateur pour le modèle standard.

Modèle de fragilité

Conditionnellement aux covariables Z, la population n'est pas forcément homogène. Soit Z' une covariable important non observée (environnement, génétique d'un population). Le modèle de Cox en incluant cette covariable est :

$$\lambda(t \mid Z, Z_0) = \lambda_0(t) \exp(\beta^\top Z + \beta'^\top Z')$$

Mais comme on ne l'observe pas, on considère que $\exp(\beta'^\top Z')$ est une variable aléatoire noté ω.
On peut par exemple attribuer une variable aléatoire w_i à différentes région du monde. Le risque instantané pour le j-ième patient du i-ième groupe est :

$$\lambda_{ij}(t \mid Z_{ij}) = \lambda_0(t)\omega_i \exp(\beta^\top Z_{ij}).$$

Les ω_i sont iid . En général :

$$\mathbb{E}(\omega) = 1 \text{ et } Var(\omega_i) = \theta.$$

Algorithme EM pour estimer les paramètres.

Exercice :

On considère le modèle de Cox :

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta^\top Z),$$

avec le risque de base instantané correspondant à une durée de survie modélisée par une loi de Weibull $\mathcal{W}(\theta, \nu)$, $\theta > 0, \nu > 0$:

$$X \sim \mathcal{W}(\theta, \nu), \;\; f_X(t|\theta, \nu) = \nu(\frac{1}{\theta})^\nu t^{\nu-1} \exp\left(-(\frac{t}{\theta})^\nu\right)$$

1. Déterminer la fonction de survie et la densité de ce modèle.
2. Comment varie le risque instantané en fonction du temps t et conditionnellement à une covariable Z.
3. Comment pourrait-on évaluer les paramètres de ce modèle ? Exprimer la fonction à maximiser en fonction d'un échantillon $(T_1, \delta_1) \ldots (T_n, \delta_n)$.

Modèle de vie accélérée

Exercice :

On considère le modèle de vie accélérée :

$$S(t \mid Z) = S_0 \left(t \exp(\beta^\top Z) \right),$$

1. Exprimer le risque instantané du modèle de vie accélérée en fonction de λ_0 le risque instantané de S_0..
2. On suppose que S_0 est la fonction de survie de la variable aléatoire $\exp(A + \epsilon)$, montrer qu'il s'agit alors d'un modèle de régression log-linéaire, càd :

$$\log(X) = A - \beta^\top Z + \epsilon.$$