# MATH 60604
# Modélisation statistique
# § 7d - Modèle à risques proportionnels de Cox

HEC Montréal
Département de sciences de la décision

# Motivation

- L'estimateur de Kaplan–Meier permet d'estimer de manière nonparamétrique la fonction de survie.
- Qu'est-ce qu'on ferait si on voulait mesurer l'effet de variables explicatives $X_1, \ldots, X_p$ sur la survie?
  - avec des variables catégorielles (et beaucoup de données), on pourra estimer la fonction pour chaque sous-groupe à l'aide de l'estimateur de Kaplan–Meier.
  - cette approche ne fonctionne pas si $X_j$ est continue ou le nombre d'observations par groupe est petite.

# Fonction de risque cumulative

Pour $T$ continue*, la fonction de risque cumulative est

$$H(t) = \int_0^t h(u)\mathrm{d}u = \int_0^t \frac{f(u)}{S(u)}\mathrm{d}u = -\ln\{S(t)\}$$

et donc on peut écrire la fonction de survie

$$S(t) = \exp\{-H(t)\}.$$

On peut aussi écrire la log vraisemblance en terme de la fonction de risque (cumulative)

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{\delta_i \ln h(t_i; \boldsymbol{\theta}) - H(t_i; \boldsymbol{\theta})\}$$

# Postulat de risques proportionnels

Dans le modèle à risques proportionnels, la fonction de risque est

$$h(t; \mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i\boldsymbol{\beta})$$

où

- la fonction de risque de base, $h_0(t)$ est le seul terme de droite qui varie dans le temps.
- l'hypothèse dite de risques proportionnels est que le rapport $h(t; \mathbf{x}_i)/h(t; \mathbf{x}_j)$ est constant peut importe la valeur de $t$.
- l'interprétation des effets des variables explicatives est simplifiée, parce que ces effets ne varient pas avec le temps.
- ce postulat est très restrictif et doit être validé en pratique, mais il est particulièrement commode pour les dérivations.

**Note:** il n'y a pas d'ordonnée à l'origine dans le modèle de Cox: cette dernière est incorporée dans $h_0(t)$.

# Dérivation du modèle à risques proportionnels

On considère les temps de défaillance observés $0 \leq t_1 < \cdots < t_D$, supposés uniques (pas de doublons) pour simplifier la dérivation.
La fonction de risque cumulative de base,

$$H_0(t) = \sum_{j:t_j \leq t} h_0(t_j),$$

est une fonction escalier avec des sauts uniquement aux temps de défaillance observés.
On considère

- $\mathcal{R}_j$, l'ensemble des individus à risques au temps $t_j$
- $\delta_i$, un indicateur binaire qui vaut 1 en cas de défaillance observée, et 0 si l'observation est censurée à droite.

# Fonction de vraisemblance du modèle de Cox

Soit $h_j = h_0(t_j)$ et $g_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$. La log vraisemblance est

$$\begin{aligned}\ell(h, \boldsymbol{\beta}) &= \sum_{i=1}^{n} \{\delta_i \ln\{\exp(\mathbf{x}_i\boldsymbol{\beta})h_i\} - \exp(\mathbf{x}_i\boldsymbol{\beta})H_0(t_j)\}\\ &= \sum_{i=1}^{n} \left\{\delta_i\mathbf{x}_i\boldsymbol{\beta} + \delta_i \ln h_i - h_i \sum_{j\in\mathcal{R}_i} \exp(\mathbf{x}_j\boldsymbol{\beta})\right\}\end{aligned}$$

- Puisqu'on s'intéresse principalement aux effets des variables explicatives $\mathbf{X}$, on considère les paramètres $h_1, \ldots, h_D$ comme des paramètres de nuisance.
- Si $\boldsymbol{\beta}$ est fixe, le maximum de vraisemblance de $h_i$ est $\widehat{h}_i = \delta_i / \sum_{j\in\mathcal{R}_i} \exp(\mathbf{x}_j\boldsymbol{\beta})$.
- Ce estimé est positif seulement si $\delta_i = 1$ (temps de défaillance observé).

# Vraisemblance profilée du modèle de Cox

On peut ainsi dériver la log vraisemblance profilée pour $\boldsymbol{\beta}$, à savoir

$$\ell_{\mathrm{p}}(\boldsymbol{\beta}) = \max_{h} \ell(h, \boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \ln\left(\frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{\sum_{j\in\mathcal{R}_i}\exp(\mathbf{x}_j\boldsymbol{\beta})}\right)$$

Il suffit alors de maximiser $\ell_{\mathrm{p}}(\boldsymbol{\beta})$. Même si ce modèle a un nombre de paramètres qui excède le nombre d'observations (!), $\ell_{\mathrm{p}}(\boldsymbol{\beta})$ se comporte à toute fin pratique comme une vraisemblance ordinaire.

- Erreurs-type via l'information observée.
- Tests de rapport de vraisemblance, du score ou de Wald pour les paramètres $\boldsymbol{\beta}$.

La situation est plus complexe s'il y a des doublons, mais les ajustements sont faits automatiquement par les logiciels (plusieurs options disponibles, certaines meilleurs et plus coûteuses que d'autres).

Une fois les estimateurs du maximum de vraisemblance $\widehat{\beta}$ recouvrés, on peut obtenir la fonction de risque cumulative estimée

$$\widehat{H}_0(t) = \sum_{i:t_i \leq t} \frac{\delta_i}{\sum_{j \in \mathcal{R}_i} \exp(\mathbf{x}_j \widehat{\beta})},$$

d'où l'estimé de la fonction de survie pour un individu avec covariables **x**

$$\widehat{S}(t; \mathbf{x}) = \exp\left\{-\exp(\mathbf{x}\widehat{\beta})\widehat{H}_0(t)\right\}$$

# Interprétations des paramètres

- Pour interpréter les paramètres du modèle de Cox à risques proportionnels, on peut comparer les taux de risques (modèle multiplicatif).
- Prenons deux individus qui sont presque identiques, sauf que leurs valeurs pour la variable $X_j$ diffère par une unité.
  - Pour l'individu *i* avec $X_{ij} = x_j$, la fonction de risque est

$$h(t; \mathbf{x}_i) = h_0(t) \exp(\beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_p x_p)$$

  - Pour l'individu *k* avec $X_{kj} = x_j + 1$, la fonction de risque est

$$h(t; \mathbf{x}_k) = h_0(t) \exp(\beta_1 x_1 + \cdots + \beta_j (x_j + 1) + \cdots + \beta_p x_p)$$

- Le rapport des fonctions de risque est

$$\frac{h(t; \mathbf{x}_k)}{h(t; \mathbf{x}_i)} = \exp(\beta_j)$$

# Rapport de risque

- Pour chaque augmentation d'une unité pour la variable $X_j$, la fonction de risque sera multipliée par un facteur $\exp(\beta_j)$, *ceteris paribus*.
  - Si $\exp(\beta_j) = 1$, $X_j$ n'a pas d'effet sur la fonction de risque.
  - Si $\exp(\beta_j) > 1$, le taux de risque **augmente** quand $X_j$ augmente.
    - Des valeurs plus élevées de $X_j$ correspondent à un risque plus élevé que l'événement survienne, et donc un temps de survie plus court.
  - Si $\exp(\beta_j) < 1$, le taux de risque **diminue** quand $X_j$ augmente.
    - Des valeurs plus élevées de $X_j$ correspondent à un risque moins élevé que l'événement survienne, et donc un temps de survie plus long.

# Exemple avec données `melanome`

Le fichier `melanome` contient des données de survie pour des patients atteints d'une tumeur maligne, un mélanome qui a été enlevé lors d'une opération chirurgicale. La base de données contient les variables suivantes:

- `temps`: temps de survie (en jours) depuis l'opération
- `statut`: 1 si le patient est mort, 0 si le temps est censuré
- `sexe`: sexe du patient, soit 1 pour les hommes et 0 pour les femmes
- `age`: l'âge (an années) au moment de l'opération
- `epaisseur`: l'épaisseur (en mm) de la tumeur
- `ulcere`: variable indicatrice, 1 en cas d'ulcération, 0 sinon.

# Statistiques descriptives pour données `melanome`

| Variable | Moyenne | Ec-type | Minimum | Maximum |
|---|---|---|---|---|
| temps | 2152.80 | 1122.06 | 10.00 | 5565.00 |
| age | 52.46 | 16.67 | 4.00 | 95.00 |
| annee | 1969.91 | 2.58 | 1962.00 | 1977.00 |
| epaisseur | 2.92 | 2.96 | 0.10 | 17.42 |

**Récapitulatif du nombre d'événements et de valeurs censurées**

| Total | Evénement | Censuré | Pourcentage censuré |
|---|---|---|---|
| 205 | 57 | 148 | 72.20 |

# Modèle de Cox pour données `melanome`

Le modèle de Cox à risques proportionnels est

$$h(t) = h_0(t)\exp(\beta_1\texttt{sexe} + \beta_2\texttt{age} + \beta_3\texttt{epaisseur} + \beta_4\texttt{ulcere})$$

On peut ajuster ce modèle dans SAS avec la procédure `phreg`:

**Code SAS pour le modèle à risque proportionnel**

```
proc phreg data=modstat.melanome;
model temps*statut(0) = sexe age epaisseur ulcere / ties=exact;
run;
```

# Tests basés sur la vraisemblance

| Statistique d'ajustement du modèle | | |
|---|---|---|
| **Critère** | **Sans covariables** | **Avec covariables** |
| **-2 LOG L** | 167.488 | 163.041 |
| **AIC** | 167.488 | 165.041 |
| **SBC** | 167.488 | 166.219 |

| Test de l'hypothèse nulle globale : BETA=0 | | | |
|---|---|---|---|
| **Test** | **khi-2** | **DDL** | **Pr > khi-2** |
| **Rapport de vrais** | 41.6195 | 4 | <.0001 |
| **Score** | 46.6689 | 4 | <.0001 |
| **Wald** | 39.4154 | 4 | <.0001 |

La sortie inclut la valeur de la log vraisemblance avec et sans variables explicatives, et les tests usuels pour $\mathscr{H}_0 : \boldsymbol{\beta} = \mathbf{0}_p$ versus $\mathscr{H}_a : \boldsymbol{\beta} \neq \mathbf{0}_p$.

# Coefficients estimés du modèle de Cox

**Analyse des valeurs estimées du maximum de vraisemblance**

| Paramètre | DDL | Valeur estimée des paramètres | Erreur type | khi-2 | Pr > khi-2 | Rapport de risque |
|---|---|---|---|---|---|---|
| **sexe** | 1 | 0.43282 | 0.26741 | 2.6197 | 0.1055 | 1.542 |
| **age** | 1 | 0.01220 | 0.00830 | 2.1616 | 0.1415 | 1.012 |
| **epaisseur** | 1 | 0.10895 | 0.03773 | 8.3362 | 0.0039 | 1.115 |
| **ulcere** | 1 | 1.16448 | 0.30975 | 14.1330 | 0.0002 | 3.204 |

# Interprétation

- Pour la variable `sexe`, $\exp(\widehat{\beta}_1) = 1.542$ représente le rapport de risque entre un homme et une femme du même âge, avec la même épaisseur de tumeur et le même état d'ulcération. Ainsi, le taux de risque pour les hommes est 1.542 fois celui pour les femmes, lorsque toutes les autres variables restent inchangées.
- Pour la variable `epaisseur`, $\exp(\widehat{\beta}_3) = 1.115$. Pour chaque augmentation de 1mm de l'épaisseur de la tumeur, le taux de risque augmente d'un facteur de 1.115 (ou 11.5%), lorsque toutes les autres variables restent inchangées.