# Analyse de survie

## Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

# Modèle à risques proportionnels de Cox

Le **modèle à risques proportionnels de Cox** pour $\mathbf{X}$ au temps $t$ est

$$h(t; \mathbf{X}) = h_0(t) \exp(\beta_1 \mathsf{X}_1 + \cdots + \beta_p \mathsf{X}_p),$$

où $h_0(t)$ est la fonction de risque de base qui remplace l'ordonnée à l'origine.

- Postulat de risques proportionnels: le rapport de risque pour deux observations ne varie pas en fonction du temps $t$.

# Postulat de risques proportionnels

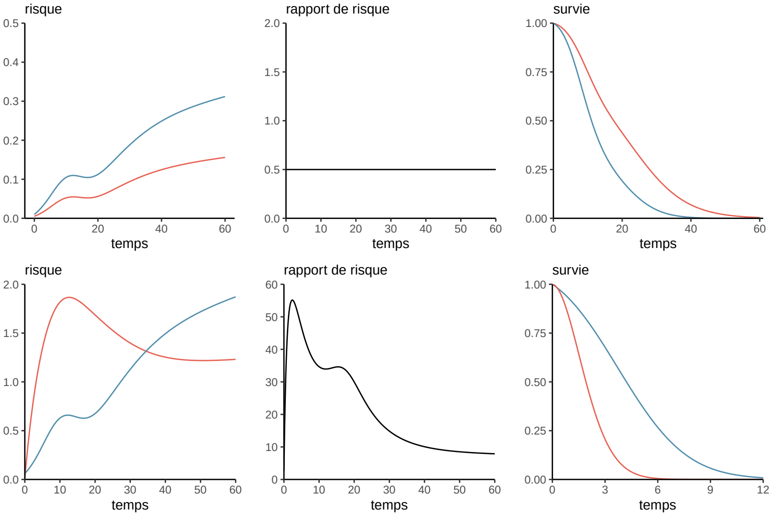



Figure 1: Courbes de risques proportionnelles (panneau supérieur) et non proportionnelles (panneau inférieur).

# Absence de proportionnalité et stratification

On peut modéliser la non proportionnalité des risques par la **stratification** pour une variable catégorielle $Z = 1, \ldots, K$.

Supposons que l'effet de $Z$ sur le risque varie dans le temps.

On écrit alors

$$h(t; \mathbf{X}, Z = k) = h_k(t) \exp(\beta_1 \mathrm{X}_1 + \cdots + \beta_p \mathrm{X}_p),$$

où $h_k(t)$ est la fonction de risque de base pour $Z = k$.

Dans ce modèle

- On suppose que l'effet des variables explicatives **X** est le même peut importe la valeur de $Z$.
- L'effet de $Z = k$ vs $Z = j$ pour un même ensemble de variables explicatives **X** est $h_k(t)/h_j(t)$, qui dépend du temps.

# Stratification

- **Avantage**: on peut modéliser n'importe quel changement du risque en fonction de $Z$.
- **Désavantage**: on perd la variable explicative $Z$, donc on ne peut tester son effet (pas de coefficient)... on peut résumer l'information pour la variable $Z$ en calculant par exemple les différences de survie à des temps donnés.
- **Désavantage**: la fonction de risque est estimée pour chaque sous-groupe de $Z$ (plus faible taille d'échantillon).

Idéalement, utiliser la stratification avec des variables secondaires ou de contrôles.

# Modèle de Cox avec stratification dans **R**

```r
library(survival)
data(survie1, package = "hecmulti")
# Stratification par service
cox_strat <- coxph(
  Surv(temps, 1-censure) ~ age + sexe + strata(service),
  data = survie1)
# Décompte par service
with(survie1, table(service))
# Coefficients
summary(cox_strat)
```

# Sorties

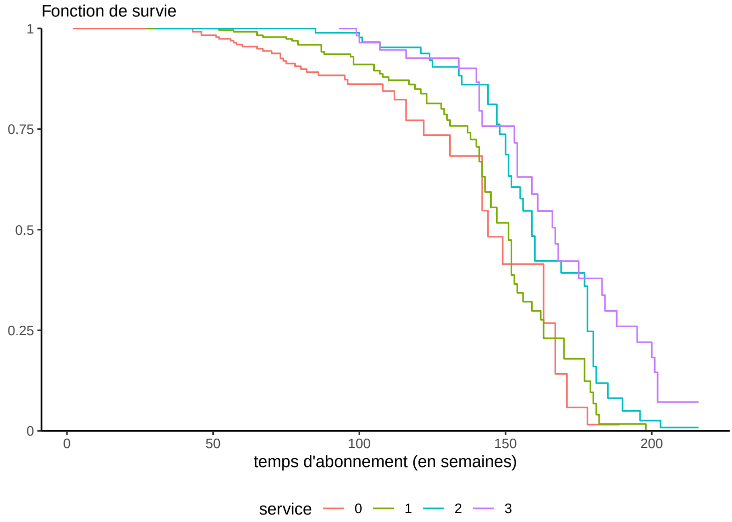
Table 1: Décompte du nombre d'observations par service.

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | 197 | 179 | 78 | 46 |

Table 2: Rapport de risques pour un modèle de Cox stratifié par service.

| terme | exp(coef) | borne inf. | borne sup. |
|---|---|---|---|
| age | 0.96 | 0.94 | 0.97 |
| sexe | 0.61 | 0.44 | 0.85 |

# Courbes de survie du modèle stratifié



# Évolution temporelle de variables explicatives

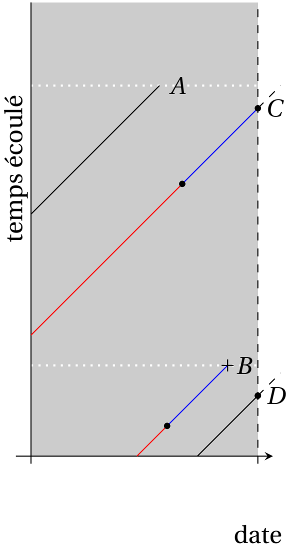
On considère une extension du modèle de Cox qui permet d'inclure des variables explicatives dont la valeur change dans le temps.

Supposons que la variable $\mathsf{X}_1$ change au fil du temps et que les autres demeurent fixes, tel que

$$h(t;x) = h_0(t)\,\mathsf{exp}\{\beta_1 \mathsf{x}_1(t) + \cdots + \beta_p \mathsf{x}_p\},$$

où $\mathsf{x}_1(t)$ indique que la valeur de $\mathsf{X}_1$ dépend du temps $t$.

# Illustration




Pour ajuster le modèle, on peut casser la contribution d'une observation en segments: considérons un seul changement survenant au temps $t_c$.

- pour le premier segment, on enregistre $t_c$ comme valeur maximale (censure à droite)
- pour la deuxième portion, l'observation est tronquée à gauche à partir de $t_c$.

# Exemple

Supposons qu'il y a eu au plus un changement dans la variable `service`. On doit formatter la base de données comme suit.

Table 3: Aperçu des cinq premières observations de la base de données survie3.

| id | debut | fin | evenement | age | sexe | region | service |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 130 | 0 | 48 | 1 | 3 | 2 |
| 1 | 130 | 178 | 0 | 48 | 1 | 3 | 1 |
| 2 | 0 | 159 | 0 | 31 | 1 | 3 | 2 |
| 3 | 0 | 110 | 1 | 36 | 1 | 4 | 0 |
| 4 | 0 | 109 | 1 | 30 | 0 | 2 | 0 |
| 5 | 0 | 78 | 0 | 22 | 0 | 5 | 1 |
| 5 | 78 | 108 | 1 | 22 | 0 | 5 | 0 |

Tout intervalle autre que terminal pour un individu est traité comme de la censure à droite.

# Code pour variables explicatives variables

```r
data(survie3, package = "hecmulti")
cox4 <- coxph(Surv(time = debut,
                   time2 = fin,
                   event = evenement) ~
                age + sexe + region + service,
              data = survie3)
```

Puisque c'est la valeur d'une variable qui varie dans le temps et non pas son effet, on a l'interprétation usuelle.

# Modèle à risques compétitifs

Parfois, la raison pour laquelle un individu quitte l'état étudié peut avoir un intérêt en soi.

Pour le temps de service d'un employé, on veut faire la distinction entre

- une démission
- un renvoi
- la retraite

# Modèle à risques compétitifs

Supposons qu'il y a $K$ manières possibles que l'événement survienne; on spécifie une fonction de risque pour chaque manière,

$$\begin{aligned} h_1(t;\mathbf{X}) &= h_{01}(t)\exp(\beta_{11}\mathsf{X}_1 + \cdots + \beta_{p1}\mathsf{X}_p) \\ &\vdots \\ h_K(t;\mathbf{X}) &= h_{0K}(t)\exp(\beta_{1K}\mathsf{X}_1 + \cdots + \beta_{pK}\mathsf{X}_p) \end{aligned}$$

Notez que les coefficients sont différents d'une équation à l'autre.

# Exemple

Supposons que nous avons trois causes possibles pour la perte d'un client. La variable `censure` dans le fichier `survie4` vaut:

- 1 si l'individu est toujours abonné à notre service
- 2 désabonnement pour aller chez le compétiteur A
- 3 désabonnement pour aller chez le compétiteur B
- 4 désabonnement parce qu'il n'a plus besoin de cellulaire.

On peut estimer les paramètres de chaque équation du modèle de Cox séparément sans perte de précision.

Deux options: ajuster chaque modèle séparément en traitant tout événement autre que celui d'intérêt comme de la censure à droite.

```
# Rappel pour "event":
#  - 1 (TRUE) pour observation,
#  - 0 (FALSE) pour censure à droite
data(survie4, package = "hecmulti")
rc_cox_A <- coxph(Surv(time = temps,
                  event = censure == 2) ~
               age + sexe + service,
             data = survie4)
```

Les observations avec des valeurs pour `censure` de 1, 3 ou 4 sont traitées comme des cas de censure à droite (l'événement quitter pour compétiteur $A$ n'est pas survenu).

# Interprétation des coefficients

Attention, l'interprétation dépend maintenant de l'événement étudié.

| terme | exp(coef) | borne inf. | borne sup. |
|---|---:|---:|---:|
| age | 0.96 | 0.94 | 0.974 |
| sexe | 0.48 | 0.35 | 0.673 |
| service1 | 0.38 | 0.27 | 0.535 |
| service2 | 0.19 | 0.11 | 0.308 |
| service3 | 0.11 | 0.05 | 0.220 |

Selo le modèle, le risque *de quitter pour aller chez le compétiteur A* d'une femme est 0.48 fois celui d'un homme.

# Code **R** pour risque compétitif (2)

Créer une variable avec identifiant `1:n` (si une seule transition est possible).

Passer la variable état comme **facteur**, avec transition depuis catégorie de référence (ici `censure=1`, soit abonné)

```r
surv4 <- survie4 |>
  dplyr::mutate(id = seq_len(dplyr::n()))
rc_cox <- coxph(Surv(time = temps,
             event = factor(censure)) ~
               sexe + age + service,
             data = surv4,
             id = id)
```

# Code **R** pour risques compétitifs (Kaplan–Meier)

Modèle avec transition d'un état de base (abonné) vers un état absorbant (désabonnement, soit chez compétiteur A ou B ou abandon du cellulaire).

Ajuster le modèle multi-état avec un **facteur** pour l'événement, où la catégorie de référence est abonnement (`censure=1`).

Surtout, ne pas estimer les courbes séparément!

```r
data(survie4, package = "hecmulti")
rc_km <- survfit(Surv(time = temps,
             event = factor(censure)) ~ 1, #facteur
             data = survie4)
```

# Graphiques

Les représentations graphiques donnent la probabilité d'être dans une situation en fonction du temps (ici avec la catégorie de référence abonnement).

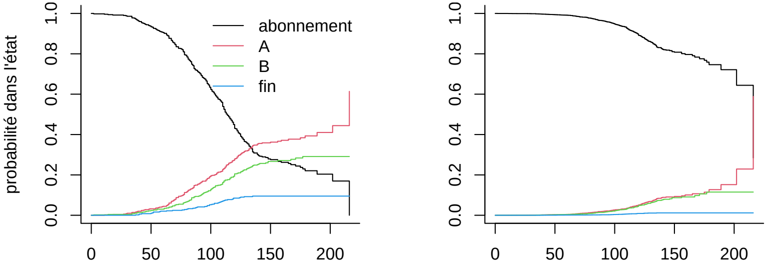



Figure 2: Probabilité d'événement sans variable explicative (Kaplan-Meier, gauche) et avec âge, service et sexe (modèle de Cox, droite).

# Risques non proportionnels

Si le postulat de risques proportionnels n'est pas validé, l'effet d'au moins une des variables explicatives dépend du temps.

On peut considérer une modification du modèle de Cox qui inclut une interaction avec le temps, par exemple

$$h(t, \mathbf{X}, Z) = h_0(t) \exp\{\beta_Z Z(t)\} \exp(\mathbf{X}\beta)$$

si l'effet de la variable explicative $Z$ — ou la variable elle même – varie en fonction du temps.

# Exemple 1 - augmentation de l'âge

L'âge (en années) augmente à mesure que le temps d'abonnement (en semaines) passe, d'où $\texttt{age}(t) = \texttt{age} + t/52$.

```
cox_np <- survival::coxph(
    Surv(temps, censure) ~
     tt(age) + sexe + service,
     data = survie1,
     tt = function(x, t, ...){x + t/52})
summary(cox_np)
```

On spécifie avec l'option `tt()` dans la formule la variable qui change dans le temps et par la suite la nature de l'interaction temporelle avec l'argument `tt`.

La variable qui dépend du temps doit être créée à l'intérieur de l'appel à `coxph`.

# Exemple 2 - effet de service croissant

Il faut transformer les variables catégorielles en indicateurs binaires.

Supposons que l'impact du nombre de services varie comme suit,

$$\begin{aligned} h(t, \text{age}, \text{sexe}, \texttt{service = i}) \\ &= h_0(t)\exp(\beta_{\texttt{sexe}}\texttt{sexe} + \beta_{\texttt{age}}\texttt{age} + \beta_{\texttt{service}_i} + \beta_{\texttt{service}_i * t}t). \end{aligned}$$

# Code **R** pour service

```r
# Créer variables binaires par service
service_bin <- model.matrix(~service,
                            data = survie1)[,-1]
# enlever ordonnée à l'origine (première colonne de uns)
# Concaténer par colonne
survie1m <- cbind(survie1, service_bin)
cox_np <- survival::coxph(
    Surv(temps, censure) ~
     age + sexe + service +
      tt(service1) + tt(service2) + tt(service3),
     data = survie1m,
     tt = function(x, t, ...){t * x})
```

# Coefficients et tests

Table 4: Rapport de risque et intervalles de confiance à niveau 95% pour le modèle à risques non proportionnels (interaction linéaire entre temps et service).

| terme | exp(coef) | test de Wald | valeur-p |
|---|---|---|---|
| age | 0.953 | -7.368 | <0.001 |
| sexe | 0.543 | -5.241 | <0.001 |
| service1 | 0.144 | -4.293 | <0.001 |
| service2 | 0.072 | -3.762 | <0.001 |
| service3 | 0.010 | -4.139 | <0.001 |
| tt(service1) | 1.010 | 2.173 | 0.0298 |
| tt(service2) | 1.010 | 1.584 | 0.1132 |
| tt(service3) | 1.023 | 2.476 | 0.0133 |

# Interprétation des résultats

- Les coefficients pour l'interaction avec $t$ sont petits parce que la plage de $t$ (0 à 200 semaines) est énorme.
- Les coefficients sont positifs: le risque augmente avec le temps. L'impact des rabais pour services multiples diminue avec le temps.
- Deux des termes d'interaction sont significatifs à niveau 5% (statistiques de Wald $Z$ de 2.173, 1.584 et 2.476 et valeurs-$p$ correspondantes de 0.03, 0.113 et 0.013).