Introduction à l'analyse de survie

Analyse de Kaplan-Meier, modélisation du risque

Paul MINCHELLA

Octobre 2025

Master MIASHS

Table des matières

Objectifs pédagogiques et compétences attendues

Ce cours a pour objectif de doter les étudiants d'une compréhension rigoureuse, mais également pratique, des modèles de survie et de leur mise en œuvre statistique. À l'issue du module, ils doivent être capables d'analyser, d'interpréter et de comparer différents modèles, selon les caractéristiques des données disponibles.

1. Compréhension du contexte et des motivations

Les méthodes de régression ou de classification classiques ne sont pas adaptées aux données de survie, car elles supposent que la variable à expliquer est toujours complètement observée. Or, dans le cadre de la survie, certaines observations sont censurées : on ne connaît pas le temps exact de survenue de l'évènement, seulement qu'il dépasse une certaine durée. Exclure ces individus reviendrait à perdre une information partielle mais précieuse, introduisant un biais dans l'analyse. De plus, le risque d'évènement évolue dans le temps et ne peut pas être traité comme une simple variable fixe. L'analyse de survie vise donc à modéliser la probabilité qu'un individu « survive » au-delà d'un instant donné, en tenant compte à la fois des évènements observés et des censures, ce que les estimateurs de Kaplan–Meier et le modèle de Cox permettent de faire rigoureusement.

- Identifier les situations où une **analyse de survie** est pertinente : temps avant la survenue d'un évènement (décès, récidive, panne, défaut, etc.), souvent en présence de **données censurées**.
- Expliquer pourquoi une approche classique de régression ou de classification n'est pas adaptée à ce type de données (non-observation de certains temps d'évènements, asymétrie d'information, dépendance au temps).
- Comprendre les notions fondamentales : fonction de survie $\mathcal{S}(t)$, fonction de risque $h(t)$, et relation entre les deux.

2. Maîtrise des méthodes descriptives : Kaplan–Meier

- Être capable, à partir d'un jeu de données, de :
  1. estimer la courbe de survie empirique par l'estimateur de Kaplan–Meier ;
  2. tracer la courbe $\widehat{\mathcal{S}}_{\mathrm{KM}}(t)$ avec ses intervalles de confiance (Greenwood) ;
  3. interpréter graphiquement les différences entre groupes.
- Appliquer et interpréter le **test du log-rank** pour comparer deux (ou plusieurs) fonctions de survie, en explicitant les hypothèses sous-jacentes (censure non-informative, indépendance des sujets, risques proportionnels).

3. Modélisation à covariables : modèle de Cox

- Construire un **modèle de Cox proportionnel** à partir de covariables explicatives.
- Comprendre que le modèle est un *modèle à risque*, dont l'objectif principal est de décrire l'influence des covariables sur le risque instantané, plutôt que de prédire le temps exact T de l'évènement.

— Savoir interpréter les coefficients estimés :

$$\widehat{\beta}_j > 0 \Rightarrow X_j \text{ augmente le risque}, \qquad \widehat{\beta}_j < 0 \Rightarrow X_j \text{ a un effet protecteur.}$$

— Construire et interpréter les **scores de risque individuels** $\eta_i = \mathbb{X}_i^\top \widehat{\boldsymbol{\beta}}$, puis réaliser une **stratification du risque** (groupes à faible, moyen ou haut risque).

— Relier cette stratification à des décisions thérapeutiques hypothétiques ou à des stratégies de suivi différenciées.

— Évaluer la qualité du modèle via :

  — des métriques de **discrimination** (C-index, AUC(t)) ;

  — des métriques de **calibration** (Brier score, calibration plot).

4. Lecture critique et comparaison des approches

Aspect	**Kaplan–Meier**	**Modèle de Cox**
Nature du modèle	Non paramétrique (aucune hypothèse sur la forme de $\mathcal{S}(t)$)	Semi-paramétrique : sépare $h_0(t)$ et $\exp(\mathbb{X}^\top \beta)$
Utilisation des covariables	Pas de covariable explicative	Prend en compte les covariables explicatives
Objectif principal	Décrire la survie empirique d'un groupe	Quantifier l'effet des covariables sur le risque instantané
Interprétation	Lecture graphique (comparaison entre groupes)	Interprétation via les hazard ratios $\exp(\beta_j)$
Avantage principal	Simple, robuste, intuitif	Flexible, permet la stratification et la prédiction relative
Limites principales	Ne permet pas d'ajuster sur plusieurs variables	Hypothèse de risques proportionnels, interprétation plus abstraite

Table 1 – Comparaison entre Kaplan–Meier et le modèle de Cox.

5. Compétences attendues

À l'issue du cours, les étudiants devront savoir :

— identifier le modèle adapté au contexte de données (KM vs Cox) ;

— implémenter les méthodes sous `Python` ou `R` (avec packages `lifelines`, `sksurv`, `survival`) ;

— interpréter correctement les sorties du modèle et justifier les décisions prises ;

— présenter de manière claire et argumentée une analyse de survie complète : estimation, tests, interprétation, validation.

L'objectif ultime n'est pas seulement d'appliquer un modèle, mais de comprendre le *raisonnement statistique* qui relie la donnée observée, la fonction de survie estimée et la prise de décision fondée sur le risque.

Chapitre 1

Notions fondamentales : variable aléatoire, trajectoire et loi

Ce chapitre introduit les notions fondamentales de probabilités nécessaires pour formaliser les modèles de survie. Avant d'aborder la modélisation du risque, il est important de comprendre ce qu'est une variable aléatoire, ce que représente une observation (ou trajectoire), et comment la loi de probabilité décrit la répartition du hasard. Ces concepts serviront de socle à toute la démarche statistique présentée dans les chapitres suivants.

1.1 Variable aléatoire : du hasard à la quantité mesurable

Définition 1.1.1 (Variable aléatoire)**.** *Une* variable aléatoire *est une quantité numérique dont la valeur dépend du hasard. Autrement dit, c'est une application qui associe à chaque issue possible d'une expérience aléatoire un nombre réel.*

De manière formelle, on note :

$$T : \Omega \longrightarrow \mathbb{R}$$
$$\omega \longmapsto T(\omega)$$

où :

- Ω *désigne l'ensemble des issues possibles de l'expérience (appelé* univers des possibles*) ;*
- $\omega \in \Omega$ *est une issue particulière (par exemple, un individu précis ou un scénario expérimental) ;*
- $T(\omega)$ *est la valeur numérique observée pour cette issue (par exemple, le temps jusqu'à la panne de cet individu).*

Ainsi, T transforme le résultat aléatoire ω en une valeur réelle mesurable et interprétable.

Le symbole Ω représente le « monde des possibles » : toutes les situations qui pourraient se produire. La variable aléatoire T est une *fonction du hasard* : avant de tirer une issue ω, sa valeur est inconnue. Une fois l'expérience réalisée (une trajectoire observée), $T(\omega)$ devient une valeur concrète, celle que l'on observe dans les données. En analyse de survie, T désigne le temps aléatoire jusqu'à l'évènement d'intérêt (décès, rechute, défaillance, etc.), et chaque sujet de la cohorte correspond à une issue ω_i.

Exemple. Si l'on observe le temps avant la panne d'un appareil, la variable aléatoire T associe à chaque appareil la durée (en jours, mois, etc.) avant sa défaillance. Avant de réaliser l'expérience, la valeur de T est inconnue, mais elle prend certaines valeurs possibles selon un mécanisme probabiliste. L'idée clé est que T n'est pas une valeur fixe, mais une variable dont le résultat est incertain. Elle modélise le hasard sous forme numérique : durée, taille, score, nombre d'essais, etc.

Après avoir introduit le concept de variable aléatoire, voyons maintenant ce que représente *une observation* concrète de cette variable, c'est-à-dire une trajectoire.

1.2 Trajectoire ou observation : la réalisation concrète du hasard

Définition 1.2.1 (Trajectoire ou observation)**.** *Une* trajectoire *(ou* réalisation*) d'une variable aléatoire correspond à la valeur effectivement observée de cette variable pour une issue particulière du hasard.*

Formellement, si $T : \Omega \to \mathbb{R}$ est une variable aléatoire, alors pour une issue donnée $\omega \in \Omega$, la quantité

$$T(\omega)$$

désigne la trajectoire *(ou la* valeur réalisée*) de T associée à cette issue ω.*

Autrement dit, $T(\omega)$ est la valeur concrète que prend la variable aléatoire T lorsque l'expérience produit le résultat ω.

Chaque issue ω du monde des possibles Ω correspond à un individu, un appareil ou un scénario expérimental particulier. Ainsi, lorsque l'on observe un échantillon de n sujets, on observe en réalité n trajectoires :

$$T(\omega_1),\ T(\omega_2),\ \ldots,\ T(\omega_n),$$

que l'on note souvent $T_1, T_2, \ldots, T_n$.

Par exemple, si T représente le temps avant la défaillance d'un appareil, une observation $T_i = 7.3$ signifie que, pour le i-ème appareil (correspondant à l'issue ω_i), la panne est survenue au bout de 7,3 jours. Chaque sujet observé fournit donc une trajectoire particulière du phénomène aléatoire global modélisé par T.

La variable aléatoire représente le *modèle abstrait du hasard*, tandis que la trajectoire correspond à une *donnée concrète*. Autrement dit : T est un concept théorique, tandis que T_i est une observation dans un jeu de données. En statistique, on observe plusieurs réalisations $T_1, T_2, \ldots, T_n$ d'une même variable aléatoire T pour en estimer les caractéristiques (moyenne, médiane, distribution...).

Nous savons maintenant distinguer la variable aléatoire (concept théorique) de son observation (valeur empirique). Il reste à comprendre comment la *loi de probabilité* décrit mathématiquement la manière dont ces valeurs se répartissent.

1.3 Loi de probabilité : description du comportement du hasard

Définition 1.3.1 (Loi de probabilité d'une variable aléatoire réelle)**.** *La* loi *(ou* distribution*) d'une variable aléatoire décrit la manière dont ses valeurs possibles se répartissent dans l'espace des nombres réels, avec leurs probabilités associées.*

Pour une variable aléatoire continue T, cette loi est décrite par :

- *une* ***fonction de répartition*** *$F(t) = \mathbb{P}(T \leq t)$, qui donne la probabilité que T prenne une valeur inférieure ou égale à t ;*
- *une* ***fonction de survie*** *$\mathcal{S}(t) = 1 - F(t) = \mathbb{P}(T > t)$, utile en analyse de survie ;*
- *une* ***densité de probabilité*** *$f(t)$ telle que $F'(t) = f(t)$ pour presque tout t.*

La loi d'une variable aléatoire est au hasard ce que la *carte de répartition* est à un territoire : elle indique quelles valeurs sont les plus ou les moins probables. Dans l'exemple de la survie, la loi de T décrit la distribution des durées de vie possibles dans la population : certaines pannes arrivent très tôt, d'autres très tard, et cette répartition est caractérisée par $\mathcal{S}$, f ou h.

Les lois usuelles que nous étudierons (exponentielle, Weibull, log-normale, Cox, etc.) spécifient des formes particulières de $\mathcal{S}$ et h.

1.4 Principe du maximum de vraisemblance

1.4.1 Motivations et intuition

L'un des fondements de l'inférence statistique consiste à *estimer des paramètres inconnus* à partir d'un échantillon observé. L'idée du **maximum de vraisemblance** (noté *MLE* pour *Maximum Likelihood Estimation*) est d'une simplicité conceptuelle remarquable : on cherche la valeur du paramètre qui rend les données observées les plus ≪ probables ≫.

Autrement dit, parmi tous les modèles possibles, on choisit celui pour lequel les observations réellement mesurées auraient eu la plus grande chance de se produire. Cette idée s'apparente à un *principe de cohérence empirique* : le modèle doit expliquer au mieux ce que l'on a observé.

Exemple 1.4.1 (Intuition)**.** *Si l'on pense qu'une variable Y suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, on ne connaît pas μ ni σ^2. Le principe de vraisemblance consiste à chercher les valeurs de μ et σ qui rendent l'échantillon observé $(y_1, \ldots, y_n)$ le plus probable sous cette hypothèse de normalité. L'estimation par maximum de vraisemblance choisira donc les paramètres $(\hat{\mu}, \hat{\sigma})$ qui maximisent la probabilité conjointe des données observées.*

Le principe du maximum de vraisemblance repose sur un raisonnement ≪ inverse ≫ : on part des observations et on cherche le modèle le plus plausible pour les avoir générées. Cette approche est universelle : qu'il s'agisse de durées de vie, de comptages, de probabilités de survie, ou de scores de risque, l'idée est toujours la même - trouver les paramètres qui rendent nos données les plus vraisemblables.

1.4.2 Formalisme mathématique

Soit un échantillon aléatoire indépendant et identiquement distribué (i.i.d.)

$$Y_1, Y_2, \ldots, Y_n \sim f(y; \theta),$$

où $f(y; \theta)$ est la fonction de densité (ou de probabilité) paramétrée par un vecteur de paramètres $\theta \in \Theta \subset \mathbb{R}^p$.

Définition 1.4.1 (Fonction de vraisemblance)**.** *La* fonction de vraisemblance *est définie par :*

$$L(\theta) = \prod_{i=1}^{n} f(y_i; \theta),$$

et représente la probabilité (ou densité) d'observer les données $(y_1, \ldots, y_n)$ sous le modèle paramétré par θ.

Définition 1.4.2 (Estimateur du maximum de vraisemblance)**.** *L'estimateur du maximum de vraisemblance (EMV) est défini comme :*

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \log L(\theta).$$

Remarque 1 (Passage au log). On travaille presque toujours avec la *log-vraisemblance* :

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(y_i; \theta),$$

car le logarithme transforme le produit en somme (plus simple à manipuler), sans changer le maximum.

Propriété 1.4.1 (Condition du maximum). *Sous des conditions de régularité, l'estimateur* $\hat{\theta}_{\text{MLE}}$ *satisfait :*

$$\left.\frac{\partial \ell(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0, \quad \text{et} \quad \left.\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0,$$

c'est-à-dire que le gradient de la log-vraisemblance s'annule au point de maximum, et que la matrice hessienne y est négative définie.

1.4.3 Propriétés et intérêt statistique

Propriété 1.4.2 (Propriétés asymptotiques). *Sous des hypothèses générales, l'estimateur du maximum de vraisemblance possède des propriétés remarquables :*

- ***Consistance*** *:* $\hat{\theta}_{\text{MLE}} \to \theta^\star$ *lorsque* $n \to \infty$ *;*
- ***Normalité asymptotique*** *:*

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \mathcal{I}(\theta^\star)^{-1}\big),$$

  où $\mathcal{I}(\theta)$ *est la* matrice d'information de Fisher *qu'on définira plus tard en équation* (8.2) *;*
- ***Efficacité asymptotique*** *: parmi les estimateurs réguliers, l'EMV atteint la variance minimale possible donnée par la borne de Cramér–Rao.*

Remarque 2 (Interprétation géométrique). On peut interpréter le maximum de vraisemblance comme la recherche du point θ qui « colle » au mieux à la forme empirique de la distribution des données. En d'autres termes, on ajuste la courbe paramétrique $f(y;\theta)$ pour qu'elle passe le plus près possible de la distribution empirique observée. Cette vision géométrique est très utile pour comprendre pourquoi l'EMV tend vers le vrai paramètre lorsque n croît.

1.4.4 Lien avec l'analyse de survie

Le principe de vraisemblance est central en analyse de survie. Dans ce contexte, la fonction de vraisemblance doit tenir compte à la fois :

- des observations complètes (sujets ayant connu l'évènement, avec densité $f(t_i)$) ;
- et des observations censurées (sujets sortis de l'étude avant l'évènement, avec contribution $\mathcal{S}(t_i)$).

Ainsi, la vraisemblance totale prend souvent la forme :

$$L(\theta) = \prod_{i=1}^{n} \big[f(t_i;\theta)\big]^{\delta_i} \big[\mathcal{S}(t_i;\theta)\big]^{1-\delta_i},$$

où $\delta_i = \mathbb{1}_{\{T_i \text{ observé}\}}$ indique si l'évènement a été observé. Ce principe sera repris et adapté dans le modèle de Cox, où seule la *vraisemblance partielle* (sans la base de risque) est utilisée pour estimer les effets des covariables.

1.5 Régression Ridge : formulation, solution et interprétation

1.5.1 Matrice de design et rappel du modèle linéaire

On considère un modèle linéaire classique :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où :

- $\mathbf{y} \in \mathbb{R}^n$ est le vecteur des réponses observées ;
- $\mathbf{X} \in \mathbb{R}^{n \times p}$ est la **matrice de design** (ou **matrice des covariables**), chaque ligne correspondant à une observation et chaque colonne à une variable explicative ;
- $\boldsymbol{\beta} \in \mathbb{R}^p$ est le vecteur des coefficients inconnus à estimer ;
- $\boldsymbol{\varepsilon}$ est le vecteur d'erreurs (supposées centrées et de variance σ^2).

1.5.2 Estimateur des moindres carrés ordinaires (OLS)

L'estimateur des moindres carrés ordinaires s'obtient en minimisant la somme des carrés des résidus :

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

Sous l'hypothèse que $\mathbf{X}^\top\mathbf{X}$ est inversible, on obtient la solution fermée :

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \quad \text{et} \quad \widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}_{\mathbf{H}}\mathbf{y},$$

où $\mathbf{H}$ est appelée la **matrice des chapeaux** (*hat matrix*).

1.5.3 Régression Ridge : formulation du problème

Le modèle Ridge introduit une pénalisation quadratique sur la norme des coefficients :

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}} = \arg\min_{\boldsymbol{\beta}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right), \qquad \lambda \geq 0.$$

Le paramètre λ contrôle l'intensité de la régularisation :

- si $\lambda = 0$, on retrouve la solution OLS classique ;
- si $\lambda \to +\infty$, les coefficients sont fortement pénalisés et tendent vers 0.

1.5.4 Démonstration de la solution fermée

On cherche à annuler le gradient de la fonction objectif :

$$\mathcal{L}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2.$$

Étape 1. Calcul du gradient.

$$\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta}.$$

Étape 2. Condition du premier ordre.

$$-2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} = \mathbf{0}.$$

Étape 3. Simplification.

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{y}.$$

Étape 4. Solution explicite.

$$\boxed{\widehat{\boldsymbol{\beta}}_{\text{Ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.}$$

On en déduit :

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{Ridge}} = \underbrace{\mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top}_{\mathbf{S}_\lambda} \mathbf{y},$$

où $\mathbf{S}_\lambda$ est appelée **matrice de lissage** (*smoother matrix*) ou **matrice des chapeaux de Ridge**.

1.5.5 Interprétation spectrale et stabilisation numérique

En notant la décomposition en valeurs singulières (SVD)

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top,$$

avec $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_p)$, on obtient :

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}} = \mathbf{V}\, \text{diag}\left(\frac{\sigma_i}{\sigma_i^2 + \lambda} \right) \mathbf{U}^\top \mathbf{y}.$$

Chaque direction principale est *rétrécie* par le facteur

$$\frac{\sigma_i}{\sigma_i^2 + \lambda} \in (0, 1].$$

Ainsi :

- les directions bien informées (grandes σ_i) sont peu affectées ;
- les directions instables (petites σ_i) sont fortement atténuées, ce qui tire les coefficients correspondants vers 0.

Conséquences pratiques :

- la matrice $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ est toujours inversible (meilleur conditionnement) ;
- la variance des estimateurs est réduite ;
- l'étude est stabilisée, surtout en présence de multicolinéarité ou de données bruitées.

Remarque. Le Ridge introduit un léger biais, mais améliore souvent la *précision prédictive* du modèle hors échantillon, grâce à un meilleur compromis biais–variance.

Chapitre 2

Motivations et introduction formelle

Ce chapitre pose le cadre conceptuel et les motivations de l'analyse de survie. Nous partons d'exemples concrets (*défaillance d'un appareil*, *rechute d'un cancer*, *décès dans une cohorte*) pour introduire les objets mathématiques centraux : le temps d'évènement T, la censure C, la durée observée $Y = \min(T, C)$, l'indicateur d'évènement $D = \mathbb{1}_{\{T \leq C\}}$, la fonction de survie $\mathcal{S}$, le hasard instantané (ou *taux de risque*) h et le risque cumulé H.

Pour donner quelques exemples concrets qui motive une telle modélisation et son emploi dans la vie de tous les jours :

— Défaillance d'un appareil électronique : On suit un lot de n disques durs à partir de leur mise en service. L'évènement d'intérêt est la première défaillance irréversible. Certains disques sont encore en fonctionnement à la fin de l'étude : leur temps de défaillance est alors *censuré à droite*. Objectifs typiques : estimer la probabilité de fonctionnement au-delà d'un horizon (fiabilité), comparer des marques ou des conditions d'usage, prévoir le risque de panne à court terme pour planifier la maintenance.

— Risque de rechute d'un cancer : Pour des patient·es traités pour un cancer, l'évènement d'intérêt est la première rechute. Les sujets ne rechutant pas avant la date de fin de suivi (ou perdus de vue) sont censurés. On souhaite estimer la *survie sans rechute*, comparer des stratégies thérapeutiques et quantifier l'effet de covariables (âge, stade, biomarqueurs).

— Décès au sein d'une cohorte : Dans une cohorte populationnelle, l'évènement est le décès (toutes causes ou cause spécifique). On étudie la survie globale, l'influence de facteurs de risque, et on rapporte des quantités interprétables (médiane de survie, taux instantané de mortalité).

Ce que l'on appelle une ***cohorte***, c'est l'ensemble de sujets (individus, dispositifs, unités biologiques, etc.) suivis dans le temps selon des règles d'inclusion et de suivi explicites, afin d'observer la survenue d'un évènement d'intérêt. Chaque sujet est décrit par des covariables de base (ex. âge, sexe, caractéristiques techniques) et éventuellement des covariables qui évoluent dans le temps.

2.1 Étude de survie : définitions opérationnelles

Définition 2.1.1 (Étude de survie)**.** *Une* étude de survie *spécifie les éléments suivants :*

(i) une ***population source*** *;*

(ii) des ***critères d'inclusion et d'exclusion*** *;*

(iii) une ***origine du temps*** *(ex. mise en service, diagnostic, randomisation) ;*

(iv) un ***évènement d'intérêt*** *(ex. défaillance, rechute, décès) ;*

(v) une ***période d'observation*** $[t_0, t_1]$ *;*

(vi) des règles de ***censure*** *(perte de vue, fin de suivi).*

Le temps d'évènement *aléatoire est noté* T*.*

Remarque 3 (Origine du temps et horloges). Le choix de l'origine (âge, temps depuis l'inclusion, temps calendaire) est crucial : il affecte l'interprétation des paramètres et la validité des comparaisons. On discutera plus loin des horloges usuelles et de l'entrée tardive (*delayed entry*).

2.2 Variables observées et notations

Définition 2.2.1 (Données élémentaires avec censure). *Soit $T \in [0, \infty]$ le temps (continu) jusqu'à l'évènement, et $C \in [0, \infty]$ le temps de censure. On* observe

$$Y = \min(T, C), \qquad D = \mathbb{1}_{\{T \leq C\}},$$

ainsi que des covariables $X \in \mathcal{X}$ (fixes ou dépendantes du temps). Pour $i = 1, \ldots, n$, on note (X_i, Y_i, D_i) les observations individuelles, supposées indépendantes ou conditionnellement indépendantes.

L'idée clé est que, dans une étude de survie, on ne connaît jamais toujours T. Si l'évènement ne s'est pas encore produit à la fin du suivi (ou si le sujet est perdu de vue), on sait seulement que $T > C$. On parle alors de *censure à droite*. La variable Y correspond donc à la durée *observée* et D indique si l'évènement a été effectivement observé ($D = 1$) ou censuré ($D = 0$).

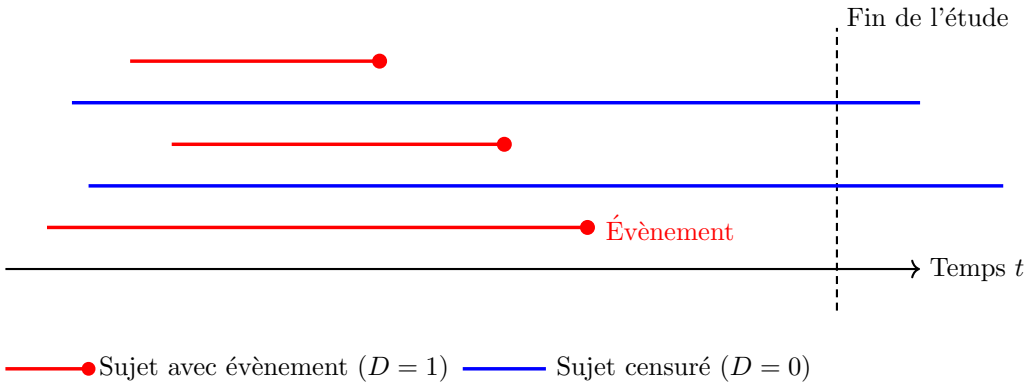

FIGURE 2.1 – Représentation schématique des temps de suivi dans une étude de survie. Les segments rouges indiquent les sujets ayant connu l'évènement ($D = 1$), tandis que les segments bleus représentent les sujets censurés ($D = 0$). La ligne verticale pointillée marque la fin de l'étude.

L'exemple ci-dessus illustre une caractéristique essentielle des données de survie : tous les sujets ne connaissent pas nécessairement l'évènement d'intérêt avant la fin de l'étude. Certains voient leur suivi interrompu prématurément (changement d'hôpital durant le suivi médical, fin d'observation, arrêt de traitement, etc.), c'est précisément la *censure*.

Ce contexte rend la modélisation classique de type « régression » inadaptée, puisqu'on ne dispose pas toujours de la variable cible complète du temps T_i pour chaque individu. Néanmoins, ces sujets censurés contiennent de l'information précieuse : le simple fait qu'ils aient survécu jusqu'à un certain temps contribue à la compréhension du phénomène étudié. Il serait donc dommage, voire statistiquement biaisé, de les exclure.

C'est précisément la force de l'analyse de survie : elle permet d'exploiter conjointement les observations complètes (évènements) et les observations incomplètes (censurées), en intégrant leur contribution respective dans une fonction de vraisemblance adaptée. Autrement dit, même un individu n'ayant pas encore connu l'évènement participe à l'inférence statistique, car on sait qu'il a « survécu » au moins jusqu'à son temps de censure.

La modélisation de la *survie* repose ainsi sur une idée simple mais puissante : apprendre à décrire, estimer et prédire le comportement du risque dans le temps, tout en respectant la nature partiellement observée des données. Ce formalisme sera la base des modèles étudiés dans les chapitres suivants – qu'ils soient paramétriques, semi-paramétriques (comme le modèle de Cox) ou modernes (modèles neuronaux de survie).

Définition 2.2.2 (Fonction de survie et fonctions associées)**.** *La* fonction de survie *de T est*

$$\mathcal{S}(t) = \mathbb{P}(T > t), \qquad t \geq 0.$$

La fonction de répartition est $F(t) = \mathbb{P}(T \leq t) = 1 - \mathcal{S}(t)$. Si T admet une densité f, alors $f(t) = F'(t)$ presque partout*.*

La fonction $\mathcal{S}(t)$ donne la probabilité de *survivre au-delà du temps t*. Elle décroît de 1 (au temps 0) vers 0 (à long terme). Graphiquement, elle illustre la proportion de sujets encore « en vie » au fil du temps. La dérivée $f(t)$ représente au contraire la *vitesse de survenue* des évènements autour de t. Ces deux notions, $\mathcal{S}$ et f, sont fondamentales et seront à la base de toutes les modélisations ultérieures.

Définition 2.2.3 (Hasard instantané et risque cumulé)**.** *Le* hasard instantané *(ou* taux de risque*) est défini par*

$$h(t) = \lim_{\mathrm{d}t \to 0} \frac{\mathbb{P}(t \leq T < t + \mathrm{d}t \mid T \geq t)}{\mathrm{d}t}, \qquad (t > 0).$$

Le risque cumulé *associé est*

$$H(t) = \int_0^t h(u)\,\mathrm{d}u.$$

Le taux de risque $h(t)$ mesure la probabilité instantanée qu'un évènement survienne à l'instant t, *sachant* que le sujet a survécu jusqu'à cet instant. Cette approche infinitésimale est en fait très intuitive : on observe le risque de survenue de l'évènement dans un intervalle de temps dt immédiatement après t, pour un sujet encore présent dans l'étude à ce moment-là.

Il s'agit donc d'un *risque conditionnel instantané*, et non d'une probabilité directe observée sur une durée finie. En intégrant ce risque instantané dans le temps, on obtient la quantité $H(t)$, qui cumule l'exposition au danger depuis le début du suivi jusqu'à l'instant t.

Comme $h(t)$ décrit le risque sur un intervalle infinitésimal, l'intégrale $\int_0^t h(u)\,\mathrm{d}u$ s'interprète comme la *somme continue* de ces micro-risques au fil du temps. Ainsi, $H(t)$ possède une interprétation concrète : il représente le *risque total accumulé* auquel un individu a été exposé entre le début de l'étude et le temps t.

Théorème 2.2.1 (Lien fondamental entre $\mathcal{S}$, h et H)**.** *Si h est localement intégrable, alors pour tout $t \geq 0$:*

$$\mathcal{S}(t) = \exp\big(-H(t)\big) = \exp\Big(-\int_0^t h(u)\,\mathrm{d}u\Big), \qquad f(t) = h(t)\,\mathcal{S}(t).$$

Preuve. Par définition du taux de risque instantané :

$$\begin{aligned}
h(t) &\underset{\mathrm{d}t\to 0}{=} \frac{\mathbb{P}\big(t \leq T < t + \mathrm{d}t \mid T > t\big)}{\mathrm{d}t} + o(1) && \text{(définition de } h(t)\text{)} \\
&\underset{\mathrm{d}t\to 0}{=} \frac{1}{\mathbb{P}(T \geq t)} \frac{\mathbb{P}\big(\{t \leq T < t + \mathrm{d}t\} \cap \{T \geq t\}\big)}{\mathrm{d}t} + o(1) && \text{(formule de Bayes)} \\
&\underset{\mathrm{d}t\to 0}{=} \frac{1}{\mathcal{S}(t)} \frac{\mathbb{P}(t \leq T < t + \mathrm{d}t)}{\mathrm{d}t} + o(1) && \text{(car } \mathcal{S}(t) = \mathbb{P}(T \geq t)\text{)} \\
&\underset{\mathrm{d}t\to 0}{=} \frac{1}{\mathcal{S}(t)} \frac{\mathbb{P}(T \leq t + \mathrm{d}t) - \mathbb{P}(T \leq t)}{\mathrm{d}t} + o(1) && \text{(propriété de } F(t)\text{)} \\
&\underset{\mathrm{d}t\to 0}{=} -\frac{1}{\mathcal{S}(t)} \frac{\mathcal{S}(t + \mathrm{d}t) - \mathcal{S}(t)}{\mathrm{d}t} + o(1) && \text{(car } F = 1 - \mathcal{S}\text{).}
\end{aligned}$$

En passant à la limite lorsque $\mathrm{d}t \to 0$, on obtient :

$$h(t) = -\frac{\mathcal{S}'(t)}{\mathcal{S}(t)} \quad \Longleftrightarrow \quad \mathcal{S}'(t) + h(t)\,\mathcal{S}(t) = 0.$$

Cette équation différentielle se résout par intégration :

$$\mathcal{S}(t) = \exp\left(-\int_0^t h(s)\,\mathrm{d}s\right).$$

On définit alors le *risque cumulé* $H(t) = \int_0^t h(s)\,\mathrm{d}s$, d'où la relation :

$$\mathcal{S}(t) = e^{-H(t)}.$$

Enfin, puisque $f(t) = F'(t) = -\mathcal{S}'(t)$, on a bien :

$$f(t) = h(t)\,\mathcal{S}(t),$$

ce qui conclut la démonstration. □

2.3 Types de censure et de troncature

Définition 2.3.1. *On définit les notions de censure à droite, à gauche, par intervalle :*

- ***Censure à droite** : on sait que $T > Y$ lorsque $D = 0$ (fin d'étude, perte de vue).*
- ***Censure à gauche** : on sait seulement que $T \leq Y$ (ex. séroconversion avant le premier test positif).*
- ***Censure par intervalle** : on sait que $T \in (L, R]$ (ex. contrôles périodiques).*

Définition 2.3.2. *On définit également la notion de troncature :*

- ***Troncature à gauche / entrée tardive** : un sujet n'entre dans la cohorte que si $T > U$ (inclusion après l'origine), ce qui conditionne le risque observable.*
- ***Troncature à droite** : seuls les sujets avec $T \leq V$ sont observables (plus rare en pratique pour la survie classique).*

Remarque 4 (Censure non-informative). Beaucoup de méthodes supposent une *censure indépendante* (*non-informative*), typiquement $T \perp\!\!\!\perp C$ (ou indépendance conditionnelle à X). Cette hypothèse garantit la cohérence d'estimateurs comme Kaplan–Meier. Sa violation nécessite des approches dédiées (IPCW, modèles conjoints, etc.).

2.4 Objets dynamiques : ensemble à risque et processus de comptage

Définition 2.4.1 (Ensemble à risque)**.** *À l'instant t, l'*ensemble à risque *est*

$$\mathcal{R}(t) = \big\{\, i \in \{1, \ldots, n\},\ Y_i \leq t \big\}.$$

Il s'agit donc de l'ensemble des sujets encore dans l'étude à l'instant t et représente la population effectivement *exposée au risque* à t.

Définition 2.4.2 (Processus de comptage N et processus d'exposition Y)**.** *On peut représenter les données par un processus de comptage*

$$N(t) = \sum_{i=1}^{n} \mathbb{1}_{\{T_i \leq t,\ D_i = 1\}}$$

et un processus d'exposition

$$Y(t) = \sum_{i=1}^{n} \mathbb{1}_{\{i \in \mathcal{R}(t)\}}.$$

Ces objets seront utiles pour les estimateurs non paramétriques et les modèles semi-paramétriques.

2.5 Quantités d'intérêt et interprétations

Définition 2.5.1. *Voici un listing non-exhaustif de quantités importantes en analyse de survie :*

- ***Médiane de survie*** *:* $\mathrm{Med}(T) = \inf\{t :\ \mathcal{S}(t) \leq 1/2\}$.
- ***Survie à un horizon*** $t^\star$ *:* $\mathcal{S}(t^\star)$.
- ***Fonction de risque conditionnelle*** *:* $t \mapsto h(t \mid \mathbb{X})$ *(si des covariables sont considérées).*
- ***Métriques en analyse de survie*** *: ex.* C-index, td-AUC, Brier Score,...

Ces différentes quantités permettent d'aborder la survie sous plusieurs angles complémentaires. La *médiane de survie* offre une mesure synthétique et robuste : elle indique le temps au-delà duquel 50% des sujets sont encore en vie.

Lorsque des covariables $\mathbb{X}$ sont introduites, la fonction $h(t \mid \mathbb{X})$ traduit la manière dont les caractéristiques individuelles modifient le risque au cours du temps. Enfin, les mesures de qualité du modèle (comme le C-index ou le Brier Score) permettent d'évaluer la *qualité prédictive* d'un modèle de survie : elles mesurent sa capacité à bien ordonner les sujets selon leur risque et à estimer correctement leurs probabilités de survie.

Ainsi, l'analyse de survie ne se limite pas à l'estimation d'une seule courbe : elle vise une compréhension fine du risque dans le temps, de ses déterminants et de la performance des modèles utilisés pour le décrire.

2.6 Hypothèses statistiques de base

Dans une cohorte convenablement conçue, on suppose généralement :

(i) des sujets indépendants (ou conditionnellement indépendants) ;

(ii) une censure **non-informative**, c'est-à-dire

$$T \perp\!\!\!\perp C \mid \mathbb{X},$$

où $\mathbb{X}$ désigne l'ensemble des covariables observées ;

(iii) un enregistrement **sans erreur** des dates clés (évènement, censure, inclusion, etc.).

Ces hypothèses constituent la base du cadre classique d'analyse de survie, mais pourront être discutées et relâchées lorsque le contexte empirique l'exigera.

Remarque 5 (Validité interne/externe et biais). Les résultats dépendent du plan d'étude (sélection, pertes de vue, délais d'inclusion). Attention aux biais de calendrier, à l'*immortal time bias* et aux confusions non mesurées. La validité externe requiert que la population source et les conditions de suivi soient comparables au contexte d'application.

Chapitre 3

Quelques lois usuelles en analyse de survie

L'objectif de ce chapitre est de présenter les principales lois de probabilité utilisées pour modéliser les temps de survie. Ces lois constituent des briques fondamentales : elles permettent d'exprimer explicitement les fonctions de survie $\mathcal{S}(t)$, de risque instantané $h(t)$ et de risque cumulé $H(t)$. Elles servent de base aux modèles paramétriques et, plus généralement, à la compréhension du comportement du risque au cours du temps.

3.1 La loi exponentielle

Définition 3.1.1 (Loi exponentielle)**.** *Une variable aléatoire T suit une* loi exponentielle *de paramètre $h > 0$, notée $T \sim \mathrm{Exp}(h)$, si sa densité est*

$$f(t) = he^{-ht}, \qquad t \geq 0.$$

Sa fonction de survie et son risque instantané sont :

$$\mathcal{S}(t) = e^{-ht}, \qquad h(t) = h.$$

Le modèle exponentiel est le plus simple possible en analyse de survie : le taux de risque $h(t)$ y est constant. Autrement dit, la probabilité instantanée d'un évènement ne dépend pas du temps écoulé. Cela revient à dire que le système ne « vieillit » pas : le risque reste identique à tout moment du suivi.

Propriété 3.1.1 (Absence de mémoire)**.** *La loi exponentielle (et son analogue discret, la loi géométrique) vérifie la propriété d'*absence de mémoire *:*

$$\mathbb{P}(T > s + t \mid T > s) = \mathbb{P}(T > t), \qquad s, t \geq 0, \quad T \sim \mathrm{Exp}(h).$$

Le caractère *sans mémoire* signifie que la probabilité de survivre encore t unités de temps ne dépend pas de la durée déjà passée. Cette propriété est intuitive dans certains contextes :

- **Appareils électroniques :** si la panne d'un composant dépend uniquement d'un choc aléatoire (et non de l'usure), le fait qu'il ait déjà fonctionné longtemps n'influence pas sa probabilité immédiate de tomber en panne.
- **Maladies chroniques :** dans certaines pathologies stables, le risque de rechute peut être considéré comme constant au cours du temps, tant qu'aucun changement d'état biologique majeur n'intervient.

Cependant, cette hypothèse est souvent trop simplificatrice : la plupart des phénomènes réels présentent un risque *variable* dans le temps ; d'où l'intérêt des lois plus flexibles comme Weibull ou log-normale.

3.2 La loi de Weibull

Définition 3.2.1 (Loi de Weibull)**.** *Une variable aléatoire T suit une* loi de Weibull *de paramètres (h, k), avec $h > 0$ et $k > 0$, notée $T \sim \text{Weibull}(k, h)$, si sa densité est :*

$$f(t) = \frac{k}{h}\left(\frac{t}{h}\right)^{k-1} e^{-(t/h)^k}, \qquad t \geq 0.$$

La fonction de survie et le risque instantané sont :

$$\mathcal{S}(t) = e^{-(t/h)^k}, \qquad h(t) = \frac{k}{h}\left(\frac{t}{h}\right)^{k-1}.$$

Remarque 6. La loi de Weibull est un *généralisateur naturel* de la loi exponentielle :

- Si $k = 1$, on retrouve l'exponentielle de paramètre h^{-1}.
- Si $k > 1$, le risque $h(t)$ *augmente avec le temps* : on modélise un phénomène de vieillissement ou d'usure.
- Si $k < 1$, le risque *diminue avec le temps* : le danger est plus élevé au début, puis décroît.

C'est pourquoi cette loi est omniprésente dans les applications industrielles (fiabilité des matériaux, durée de vie des composants) et médicales (évolution d'un risque avec le temps depuis un traitement ou une chirurgie). Elle offre un compromis idéal entre simplicité analytique et flexibilité du profil de risque.

3.3 Autres lois paramétriques utiles

Définition 3.3.1 (Loi log-normale)**.** *Une variable aléatoire T suit une loi* log-normale *si $\log(T)$ suit une loi normale de moyenne μ et d'écart type $\sigma > 0$. On note alors $T \sim \text{LogNormal}(\mu, \sigma)$, avec densité :*

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right), \qquad t > 0.$$

La log-normale modélise bien les phénomènes dont les durées résultent de la *multiplication* de plusieurs facteurs indépendants (par opposition à la somme pour la normale). Son risque $h(t)$ est souvent en forme de cloche : faible au début, maximal à un certain instant, puis décroissant. C'est un bon choix pour des maladies dont le risque de rechute est maximal dans une période critique après traitement, puis diminue.

Définition 3.3.2 (Loi log-logistique)**.** $T \sim \text{LogLogistic}(\alpha, \beta)$ *si*

$$\mathcal{S}(t) = \frac{1}{1 + (t/\alpha)^\beta}, \qquad t > 0.$$

Le risque instantané est :

$$h(t) = \frac{\frac{\beta}{\alpha}(t/\alpha)^{\beta-1}}{1 + (t/\alpha)^\beta}.$$

Cette loi est très utilisée pour sa simplicité analytique et sa capacité à générer des profils de risque non monotones (croissant puis décroissant). Elle permet également des expressions fermées pour la médiane et la fonction de survie, facilitant l'interprétation et l'estimation.

Chapitre 4

Modélisation de la fonction de survie : approche de Kaplan–Meier

La fonction de survie $\mathcal{S}(t) = \mathbb{P}(T > t)$ décrit la probabilité qu'un individu ou un système n'ait pas encore connu l'évènement d'intérêt à l'instant t. Dans les chapitres précédents, nous avons vu comment $\mathcal{S}$ s'interprète et comment certaines lois paramétriques (exponentielle, Weibull, etc.) en proposent des formes analytiques.

Cependant, dans de nombreuses situations pratiques, on ne souhaite pas imposer de forme particulière à la loi de T. On cherche alors à *estimer empiriquement* la fonction de survie à partir des données observées (Y_i, D_i), tout en tenant compte de la censure. L'estimateur de Kaplan–Meier [KM58], aussi appelé *estimateur du produit limite*, constitue la méthode standard pour cette tâche.

4.1 Principe et intuition

L'idée de Kaplan–Meier (1958) repose sur une approche simple et intuitive : on estime la probabilité de survivre jusqu'à un instant t comme le *produit* des probabilités de survivre à chacun des instants où un évènement a été observé.
En d'autres termes, au lieu de modéliser une densité $f(t)$ ou un risque $h(t)$, on s'intéresse directement à la probabilité empirique de survie :

$$\mathcal{S}(t) = \mathbb{P}(T > t).$$

Chaque fois qu'un évènement (décès, panne, rechute, etc.) est observé, la fonction de survie décroît par un facteur correspondant à la proportion d'individus ayant connu l'évènement à cet instant.
On observe un échantillon (Y_i, D_i) pour $i = 1, \ldots, n$, où :

$$Y_i = \min(T_i, C_i), \qquad D_i = \mathbb{1}_{\{T_i \leq C_i\}}.$$

On suppose les n individus indépendants et la **censure non-informative** : $T \perp\!\!\!\perp C$.

4.2 Construction de l'estimateur de Kaplan–Meier

Définition 4.2.1 (Estimateur de Kaplan–Meier)**.** *Soient $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$ les instants distincts où au moins un évènement est observé ($D_i = 1$). On note :*

- *d_j : le nombre d'évènements observés à l'instant $t_{(j)}$;*
- *n_j : le nombre d'individus encore « à risque » juste avant $t_{(j)}$, c'est-à-dire ceux tels que $Y_i \geq t_{(j)}$.*

*L'*estimateur de Kaplan–Meier *de la fonction de survie est défini par :*

$$\widehat{\mathcal{S}}_{\mathrm{KM}}(t) = \prod_{j : t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right). \tag{4.1}$$

Chaque facteur $\left(1 - \frac{d_j}{n_j}\right)$ représente la probabilité empirique de *survivre à l'instant* $t_{(j)}$ sachant qu'on était encore à risque juste avant cet instant. Le produit accumule ces probabilités au fil du temps, donnant ainsi la probabilité de n'avoir connu aucun évènement jusqu'à t.

Graphiquement, $\mathcal{S}_{\mathrm{KM}}(t)$ est une *fonction en escalier décroissante* : elle reste constante entre deux évènements observés et chute d'un cran à chaque nouvelle occurrence.

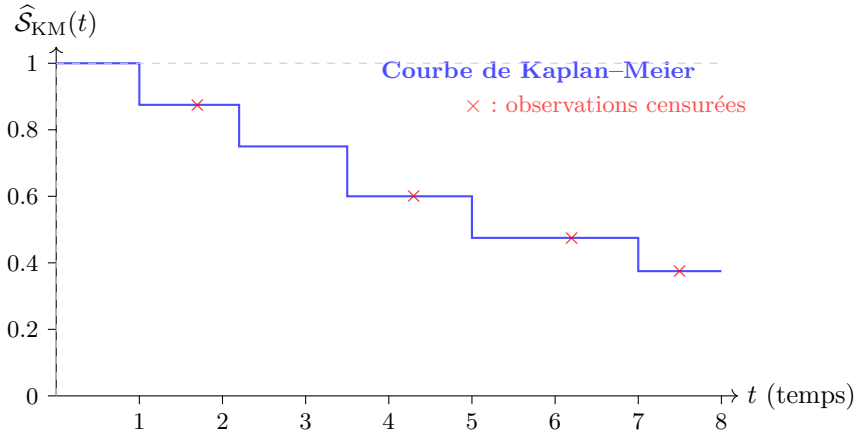

FIGURE 4.1 – Exemple de fonction de survie estimée par l'approche de Kaplan–Meier avec indicateurs de censure (croix rouges).

L'expression de Kaplan–Meier peut se lire comme un **produit cumulatif de probabilités conditionnelles de survie**. En effet, pour chaque instant $t_{(j)}$ où un évènement est observé :

$$\widehat{\mathcal{S}}_{\mathrm{KM}}(t) = \prod_{j : t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right) = \underbrace{\left(1 - \frac{d_1}{n_1}\right)}_{\text{Survie jusqu'à } t_{(1)}} \times \underbrace{\left(1 - \frac{d_2}{n_2}\right)}_{\text{Survie de } t_{(1)} \text{ à } t_{(2)}} \times \cdots \times \underbrace{\left(1 - \frac{d_j}{n_j}\right)}_{\text{Survie de } t_{(j-1)} \text{ à } t_{(j)}}.$$

Chaque facteur $\left(1 - \frac{d_j}{n_j}\right)$ représente la probabilité empirique de survivre à $t_{(j)}$, conditionnellement au fait d'être encore à risque juste avant cet instant. Le produit accumule ces probabilités élémentaires, traduisant le fait que la survie totale à l'instant t est la probabilité d'avoir « survécu à toutes les étapes précédentes ».

4.3 Propriétés fondamentales

Propriété 4.3.1 (Estimée empirique non paramétrique)**.** *L'estimateur de Kaplan–Meier est* **non paramétrique** *: il ne suppose aucune forme fonctionnelle pour la loi de survie. Il s'appuie uniquement sur les proportions observées (d_j, n_j), et incorpore naturellement la censure à droite.*

Propriété 4.3.2 (Lien avec la vraisemblance)**.** *L'estimateur de Kaplan–Meier peut être vu comme le* maximum de vraisemblance non paramétrique *sous hypothèse d'indépendance et de censure non-informative. En effet, la vraisemblance empirique des observations (Y_i, D_i) conduit, par maximisation, à l'expression du produit*

$$\hat{\mathcal{S}}(t) = \prod_{j:t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

Idée de preuve. On montre que l'expression de Kaplan–Meier maximise la vraisemblance empirique sous les hypothèses :

$$T_1, \ldots, T_n \text{ indépendants}, \quad T_i \perp\!\!\!\perp C_i, \quad \text{et censure à droite.}$$

La vraisemblance complète des observations (Y_i, D_i) s'écrit :

$$L = \prod_{i=1}^{n} \big[f(Y_i)\big]^{D_i} \big[\mathcal{S}(Y_i)\big]^{1-D_i},$$

où f et $\mathcal{S}$ désignent la densité et la fonction de survie de T.

Regroupons maintenant les observations par temps distincts d'évènements $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$. Notons d_j le nombre d'évènements à $t_{(j)}$ et n_j le nombre d'individus encore à risque juste avant $t_{(j)}$. Sous une approche non paramétrique, on considère que f (ou la loi de T) ne prend des valeurs non nulles qu'aux instants d'évènement $t_{(j)}$: autrement dit, on cherche directement à estimer les probabilités atomiques

$$p_j = \mathbb{P}(T = t_{(j)} \mid T \geq t_{(j)}).$$

La vraisemblance empirique peut alors se réécrire comme :

$$L = \prod_{j=1}^{m} p_j^{d_j} (1 - p_j)^{n_j - d_j}.$$

Cette expression correspond exactement à un produit de lois binomiales : à chaque temps $t_{(j)}$, d_j succès (évènements) sont observés parmi n_j individus à risque.

Maximiser L par rapport à p_j donne

$$\widehat{p}_j = \frac{d_j}{n_j}.$$

En réinjectant dans la définition cumulative de la survie :

$$\widehat{\mathcal{S}}(t) = \prod_{j:t_{(j)} \leq t} (1 - \widehat{p}_j) = \prod_{j:t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

On retrouve donc la formule de Kaplan–Meier comme le *maximum de vraisemblance non paramétrique* estimant la fonction de survie. □

L'estimateur de Kaplan–Meier ne sort pas de nulle part : il découle directement du principe de vraisemblance, appliqué à une loi inconnue et non paramétrée. L'idée fondamentale est la suivante : on cherche à estimer la fonction de survie $\mathcal{S}(t)$ sans supposer de forme analytique particulière (exponentielle, Weibull, etc.). Plutôt que d'imposer une densité f_θ paramétrée, on considère que la loi de T

est *discrète* sur les instants où des évènements ont été observés — autrement dit, toute l'information utile est contenue dans les couples (d_j, n_j).

Ainsi, à chaque instant $t_{(j)}$, on observe d_j évènements parmi n_j sujets encore à risque. On peut alors voir cette situation comme un petit *essai binomial* :

$$\ll \text{ parmi } n_j \text{ individus à risque, } d_j \text{ connaissent l'évènement } \gg.$$

La probabilité d'un évènement à cet instant est notée p_j. En maximisant la vraisemblance empirique de ces observations binomiales successives, on trouve naturellement

$$\widehat{p}_j = \frac{d_j}{n_j}.$$

Ce résultat a une interprétation simple : c'est la proportion empirique d'évènements parmi ceux encore exposés au risque à ce moment précis.

Enfin, la probabilité de *survivre jusqu'à* t correspond à la probabilité de n'avoir subi *aucun de ces petits évènements successifs*, d'où le produit :

$$\widehat{\mathcal{S}}(t) = \prod_{j:t_{(j)}\leq t} (1-\widehat{p}_j) = \prod_{j:t_{(j)}\leq t} \left(1 - \frac{d_j}{n_j}\right).$$

Cette structure multiplicative exprime exactement la probabilité cumulée de survivre à chaque étape du suivi.

Ainsi, l'estimateur de Kaplan–Meier apparaît à la fois comme :

- une **construction empirique intuitive**, reposant sur des proportions observées ;
- et une **solution optimale au sens de la vraisemblance**, obtenue sans aucun paramètre imposé.

Ce double statut explique à la fois la simplicité, la robustesse et la popularité de cette approche en analyse de survie.

Propriété 4.3.3 (Valeurs extrêmes et continuité)**.** *L'estimateur vérifie :*

$$\mathcal{S}_{\mathrm{KM}}(0) = 1, \qquad \lim_{t\to+\infty} \mathcal{S}_{\mathrm{KM}}(t) = \mathcal{S}_{\mathrm{KM}}(t_{(m)}),$$

c'est-à-dire qu'il reste constant après le dernier évènement observé. Il est à droite-continue et décroissante, avec des discontinuités aux instants d'évènement.

Propriété 4.3.4 (Variance et intervalle de confiance [Gre26])**.** *Une estimation classique de la variance de* $\mathcal{S}_{\mathrm{KM}}(t)$ *est donnée par la formule de Greenwood :*

$$\widehat{\mathrm{Var}}\big[\mathcal{S}_{\mathrm{KM}}(t)\big] = \mathcal{S}_{\mathrm{KM}}(t)^2 \sum_{j:t_{(j)}\leq t} \frac{d_j}{n_j(n_j-d_j)}.$$

Cette variance permet de construire un intervalle de confiance pour la fonction de survie, par exemple via une transformation logarithmique :

$$\left[\log\big(-\log\widehat{\mathcal{S}}_{\mathrm{KM}}(t)\big) \pm z_{1-\alpha/2}\, \frac{\sqrt{\widehat{\mathrm{Var}}\big[\mathcal{S}_{\mathrm{KM}}(t)\big]}}{\widehat{\mathcal{S}}_{\mathrm{KM}}(t)\,\log\widehat{\mathcal{S}}_{\mathrm{KM}}(t)}\right].$$

avec $z_{1-\alpha/2}$ *le quantile d'ordre* $1-\alpha/2$ *d'une* $\mathcal{N}(0,1)$.

4.4 Interprétation et visualisation

La courbe de Kaplan–Meier fournit une représentation simple et immédiate de la survie empirique d'une population. Chaque palier horizontal correspond à une période sans évènement, tandis que chaque chute correspond à un ou plusieurs évènements observés. Les observations censurées apparaissent souvent comme des marques (croix, tirets) sur la courbe : elles indiquent des sujets dont le suivi s'est arrêté sans évènement. Cette représentation est particulièrement utile pour :

- comparer visuellement plusieurs groupes (ex. : patients traités vs non traités) ;
- estimer la médiane de survie, définie par $\widehat{\mathrm{Med}}(T) = \inf\{t : \widehat{\mathcal{S}}(t) \leq 1/2\}$;
- évaluer rapidement la proportion d'individus encore « vivants » à un instant donné.
- interpréter directement la survie comme une probabilité : par exemple, si $\widehat{\mathcal{S}}(t) = 0.8$, cela signifie qu'à l'instant t, la probabilité estimée qu'un individu de la cohorte soit encore en vie (ou n'ait pas connu l'évènement) est de 80%.

4.5 Application : comparaison entre groupes

Définition 4.5.1 (Fonction de survie par groupe)**.** *Supposons que la population soit divisée en K groupes distincts (ex. traitements). Pour chaque groupe k, on estime une fonction de survie $\mathcal{S}_k(t)$ à l'aide de l'estimateur de Kaplan–Meier calculé sur les observations de ce groupe uniquement :*

$$\widehat{\mathcal{S}}_k(t) = \prod_{j : t_{k,(j)} \leq t} \left(1 - \frac{d_{k,j}}{n_{k,j}}\right).$$

Le test du log-rank, introduit par Mantel [Man66] et formalisé par les frères Peto [PP72], permet de comparer deux courbes de survie sous hypothèse de censure non-informative.

Définition 4.5.2 (Test du log-rank)**.** *On souhaite tester l'hypothèse selon laquelle deux groupes présentent la même fonction de survie. Les hypothèses sont :*

$$(H_0) : \quad \forall t \geq 0,\ \mathcal{S}_1(t) = \mathcal{S}_2(t),$$

$$(H_1) : \quad \exists\, t \geq 0,\ \mathcal{S}_1(t) \neq \mathcal{S}_2(t).$$

Le test du log-rank *repose sur la comparaison, à chaque instant d'évènement $t_{(j)}$, du nombre observé d'évènements dans le groupe 1 (d_{1j}) avec le nombre attendu e_{1j} sous l'hypothèse nulle H_0. On définit alors la statistique de test :*

$$Z = \frac{\sum_j (d_{1j} - e_{1j})}{\sqrt{\sum_j v_{1j}}},$$

où v_{1j} est la variance associée à la différence $(d_{1j} - e_{1j})$ sous H_0.

Sous l'hypothèse nulle H_0, la statistique

$$Z^2 \sim \chi^2_1 \quad \textit{(asymptotiquement).}$$

On en déduit la ***p-valeur*** *:*

$$\textit{p-value} = \mathbb{P}\big(\chi^2_1 \geq Z^2_{\mathrm{obs}}\big),$$

qui mesure la probabilité d'obtenir une statistique au moins aussi extrême que celle observée, si H_0 était vraie.

*La **règle de décision** est la suivante :*

$$\textit{si } p\text{-value} < \alpha, \quad \textit{on rejette } H_0,$$

au seuil de signification α (souvent fixé à 5%). Dans ce cas, on conclut à une différence significative entre les deux courbes de survie.

Remarque 7 (Intérêt pratique). Le test du log-rank est la méthode standard pour comparer des courbes de survie. Il est particulièrement robuste et non paramétrique : il ne suppose aucune forme particulière du risque. Cependant, il est surtout sensible aux différences globales de survie et peut manquer de puissance lorsque les courbes se croisent.

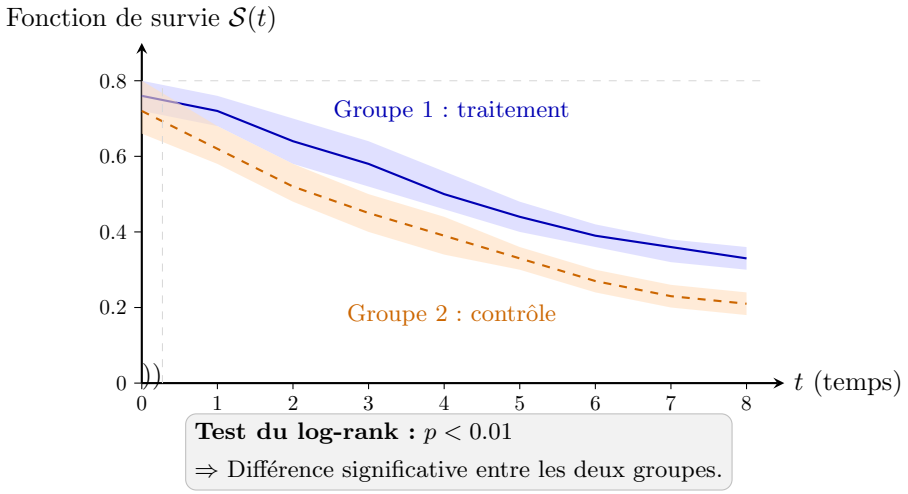

FIGURE 4.2 – Exemple illustratif d'une comparaison de deux courbes de survie estimées (Kaplan–Meier) avec intervalles de confiance à 95%. Le résultat renvoyé par le test du log-rank permet de conclure quant à l'efficacité du traitement.

4.6 Résumé et portée de l'approche Kaplan–Meier

Synthèse. L'approche de Kaplan–Meier présente plusieurs avantages majeurs :

- elle permet d'estimer la fonction de survie sans hypothèse paramétrique ;
- elle intègre naturellement la censure à droite ;
- elle fournit une estimation simple, monotone et interprétable de $\mathcal{S}(t)$;
- elle constitue la base de la plupart des comparaisons entre groupes (log-rank test) et des représentations graphiques de survie.

Cependant, elle ne permet pas de prendre en compte directement des covariables explicatives continues ou catégorielles : pour cela, il faudra recourir à des modèles de régression, tels que le modèle de Cox, étudié au chapitre suivant.

Chapitre 5

Régression linéaire généralisée

L'analyse de survie s'appuie souvent sur des modèles où la *relation entre les covariables et la réponse* n'est pas strictement linéaire. Avant d'introduire le modèle de Cox, il est utile de rappeler brièvement la structure générale des *modèles linéaires généralisés* (GLM), qui unifient les approches de régression linéaire, logistique et de Poisson.

5.1 Forme générale du modèle

Définition 5.1.1 (Modèle linéaire généralisé (GLM))**.** *Un* modèle linéaire généralisé *relie la moyenne conditionnelle d'une variable réponse Y à une combinaison linéaire de covariables $\mathbb{X} = (X^1, \ldots, X^p)$ à travers une fonction de lien g :*

$$g\big(\mathbb{E}[Y \mid X^1, \ldots, X^p]\big) = \beta_0 + \sum_{j=1}^{p} \beta_j X^j.$$

Ici :

- $\mathbb{E}[Y \mid \mathbb{X}]$ *désigne la moyenne de Y conditionnellement aux covariables $\mathbb{X}$;*
- *g est une* fonction de lien *(ex. : identité, logit, log, etc.) ;*
- *$\beta_0, \ldots, \beta_p$ sont les paramètres du modèle (ou* poids*) à estimer.*

Le terme « linéaire généralisé » signifie que la *combinaison des covariables* reste linéaire, mais que la *relation entre cette combinaison et la moyenne de Y* est modulée par une fonction de lien g. Ainsi, la linéarité s'applique dans l'espace transformé par g, ce qui permet de traiter aussi bien des réponses continues que binaires ou discrètes.

5.2 Exemples de modèles particuliers

Exemple 5.2.1 (Régression linéaire classique)**.** *Pour une variable réponse continue $Y \in \mathbb{R}$, on prend la fonction de lien identité : $g(y) = y$, d'où :*

$$\mathbb{E}[Y \mid X] = \beta_0 + \sum_{j=1}^{p} \beta_j X^j.$$

Ce modèle suppose que la variance de Y est constante et indépendante de sa moyenne.

Remarque 8. La régression linéaire ordinaire est donc un cas particulier du modèle linéaire généralisé avec $g(y) = y$. Le terme "généralisé" n'ajoute rien ici, mais il prépare à des liens plus complexes où la moyenne ne dépend pas linéairement des covariables.

Exemple 5.2.2 (Régression logistique)**.** *Lorsque Y est binaire ($Y \in \{0,1\}$), on cherche à modéliser la probabilité $p(x) = \mathbb{P}(Y = 1 \mid X = x)$. Le modèle logistique utilise le lien* logit *:*

$$g(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X^j.$$

On obtient donc :

$$p(x) = \frac{\exp(\beta_0 + \beta^\top X)}{1 + \exp(\beta_0 + \beta^\top X)}.$$

Remarque 9. Le lien logit rend la régression linéaire compatible avec une réponse bornée entre 0 et 1. Chaque coefficient β_j mesure l'effet d'unité de X^j sur le *log-odds* de succès :

$$\log\left(\frac{p}{1-p}\right).$$

Une variation de X^j d'une unité multiplie donc les *odds* par $\exp(\beta_j)$.

Exemple 5.2.3 (Régression de Poisson)**.** *Pour modéliser des comptages ($Y \in \mathbb{N}$), on suppose :*

$$Y \mid X \sim \textit{Poisson}(h(X)), \qquad \textit{avec } g(h) = \log(h).$$

Ainsi :

$$\log \mathbb{E}[Y \mid X] = \beta_0 + \sum_{j=1}^{p} \beta_j X^j.$$

Remarque 10. La régression de Poisson est très proche, conceptuellement, des modèles utilisés en analyse de survie : elle relie une intensité (ici, $h(X)$) à un prédicteur linéaire via une transformation logarithmique. Cette analogie prépare naturellement le terrain pour le modèle de Cox, où la *fonction de risque instantané* jouera un rôle analogue à $h(X)$.

5.3 Perspective vers le modèle de Cox

Propriété 5.3.1 (Principe commun aux GLM et au modèle de Cox)**.** *Dans un GLM comme dans le modèle de Cox, on cherche à relier une quantité d'intérêt (moyenne, risque, intensité) à des covariables à travers un prédicteur linéaire :*

$$\textit{Quantité d'intérêt} = g^{-1}(\beta_0 + \beta^\top X).$$

La principale différence est que, dans le modèle de Cox, la quantité étudiée n'est plus une moyenne ou une probabilité, mais un taux de risque instantané *dépendant du temps.*

Les modèles linéaires généralisés fournissent donc le cadre conceptuel dans lequel s'inscrit le modèle de Cox : on conserve la linéarité des effets des covariables, mais la fonction cible devient le *logarithme du risque instantané*. C'est cette idée – un modèle semi-paramétrique avec lien multiplicatif sur le risque – que l'on développera dans le chapitre suivant.

Chapitre 6

Le modèle de Cox : fondements et estimation

Le modèle de Cox [Cox72], proposé en 1972, est l'un des piliers de l'analyse de survie moderne. Il offre un cadre flexible permettant d'étudier l'effet de covariables sur le risque d'un évènement, sans supposer de forme particulière pour la loi du temps de survie.

Ce modèle repose sur une hypothèse dite de *séparabilité* : la dépendance au temps et celle aux covariables sont multiplicatives dans le taux de risque. Avant d'introduire cette idée, commençons par rappeler ce que signifie le terme *semi-paramétrique*.

6.1 Modèles semi-paramétriques

Définition 6.1.1 (Modèle semi-paramétrique)**.** *Un* modèle semi-paramétrique *est un modèle comportant à la fois :*

- *une composante* ***paramétrique****, de dimension finie (ici, les coefficients $\beta \in \mathbb{R}^p$) ;*
- *une composante* ***non paramétrique****, de dimension infinie (ici, la fonction de risque de base $h_0(t)$).*

Ainsi, contrairement à un modèle entièrement paramétrique (ex. : exponentiel, Weibull), le modèle semi-paramétrique ne suppose aucune forme particulière pour la distribution du temps de survie.

Dans un modèle semi-paramétrique, on cherche à capturer les effets systématiques des covariables par une structure paramétrique, tout en laissant la dynamique temporelle libre. Ce compromis combine *interprétabilité* et *flexibilité*, en particulier pour la modélisation de phénomènes de durée où la forme de la fonction de survie est souvent inconnue.

6.2 Modèle de séparabilité

Définition 6.2.1 (Modèle de Cox [Cox72])**.** *Le modèle de Cox postule une* ***séparabilité multiplicative*** *du risque instantané :*

$$h(t \mid \mathbb{X}) = h_0(t)\, \exp\big(\beta^\top \mathbb{X}\big)\,, \tag{6.1}$$

où :

- *$h_0(t)$ est la* ***fonction de risque de base*** *(non paramétrique), décrivant la dynamique temporelle commune à tous les individus ;*
- *$\exp(\beta^\top \mathbb{X})$ est un* ***facteur multiplicatif*** *décrivant l'effet des covariables sur le risque relatif.*
- *La forme linéaire $\beta^\top \mathbb{X}$, souvent noté η, s'appelle le* **score de risque**.

Cette hypothèse de "séparabilité" est analogue à celle rencontrée en physique : le risque dépend à la fois d'un *facteur temporel global* $h_0(t)$ et d'un *facteur structurel* propre à chaque individu. Autrement dit, la dynamique du temps (le "profil de dangerosité" du système) est séparée de l'effet des caractéristiques observées. Cette décomposition multiplicative rend le modèle particulièrement interprétable.

Propriété 6.2.1 (Risque relatif)**.** *Pour deux individus i et j de covariables $\mathbb{X}_i$ et $\mathbb{X}_j$, le rapport de leurs risques instantanés est :*

$$\frac{h(t \mid \mathbb{X}_i)}{h(t \mid \mathbb{X}_j)} = \exp\big(\beta^\top(\mathbb{X}_i - \mathbb{X}_j)\big),$$

indépendamment du temps t.

Cette propriété illustre l'hypothèse fondamentale du modèle de Cox : les risques relatifs entre individus sont constants dans le temps. Les courbes de survie peuvent donc se croiser verticalement (différence de niveau) mais pas horizontalement (différence de forme).

6.3 Vraisemblance partielle

Définition 6.3.1 (Vraisemblance partielle de Cox)**.** *Plutôt que de modéliser $h_0(t)$, Cox a introduit la* vraisemblance partielle *en 1975 [Cox75], qui permet d'estimer β sans spécifier $h_0(t)$. Soient $t_{(1)} < \cdots < t_{(m)}$ les instants distincts d'évènements observés. La vraisemblance partielle s'écrit :*

$$L_p(\beta) = \prod_{j=1}^{m} \frac{\exp(\beta^\top \mathbb{X}_j)}{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\beta^\top \mathbb{X}_i)}, \tag{6.2}$$

où $\mathcal{R}(t_{(j)})$ désigne l'ensemble des individus encore à risque juste avant $t_{(j)}$.

L'idée est de comparer, à chaque instant d'évènement, la probabilité que l'individu i défaillant soit celui qui subisse l'évènement parmi tous ceux encore à risque. La fonction de risque de base $h_0(t)$ disparaît dans le rapport, ce qui permet d'estimer β sans en faire d'hypothèse paramétrique. De plus, la forme de la vraisemblance partielle de Cox rappelle fortement la fonction *softmax* utilisée en apprentissage statistique. En effet, l'expression (6.2) correspond exactement à une normalisation exponentielle des scores $\beta^\top \mathbb{X}_j$, c'est-à-dire à un *softmax* calculé sur l'ensemble des individus à risque.

Ainsi, la vraisemblance partielle de Cox s'interprète comme une somme de log-vraisemblances de type *softmax*, où chaque individu à risque reçoit un score $\exp(\beta^\top \mathbb{X}_i)$, et la probabilité d'être l'évènement observé à un instant donné correspond à la normalisation de ces scores parmi tous les individus encore en compétition.

Propriété 6.3.1 (Log-vraisemblance partielle)**.** *En pratique, on maximise le logarithme de $L_p(\beta)$:*

$$\ell_p(\beta) = \sum_{j=1}^{m} \left[\beta^\top \mathbb{X}_{(j)} - \log\left(\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\beta^\top \mathbb{X}_i) \right) \right]. \tag{6.3}$$

Remarque 11 (Lien avec la vraisemblance complète). La vraisemblance partielle est issue de la vraisemblance complète du modèle de Cox après simplification des termes communs à tous les individus. Elle retient uniquement les composantes dépendantes de β et donc des effets des covariables.

Propriété 6.3.2 (Estimation du vecteur de coefficients $\widehat{\beta}$)**.** *L'estimateur du vecteur de coefficients est obtenu par maximum de vraisemblance exprimée en* (6.3) *:*

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmax}}\, \ell_p(\beta),$$

c'est-à-dire le point qui maximise la vraisemblance (ou minimise la log-vraisemblance négative).

Remarque 12 (Conditions du maximum et interprétation). L'optimum $\widehat{\beta}$ vérifie la condition du premier ordre :

$$\frac{\partial \ell_p(\beta)}{\partial \beta}\Big|_{\beta=\widehat{\beta}} = 0,$$

soit, composante par composante :

$$\sum_{j=1}^{m} \Big[\mathbb{X}_{(j)} - \frac{\sum_{i \in \mathcal{R}(t_{(j)})} \mathbb{X}_i \exp(\widehat{\beta}^\top \mathbb{X}_i)}{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\widehat{\beta}^\top \mathbb{X}_i)} \Big] = 0.$$

Cette équation exprime un équilibre entre les covariables observées chez les sujets ayant connu l'évènement et leur moyenne pondérée parmi tous les individus encore à risque au même instant.

Intuitivement, le modèle cherche un vecteur $\widehat{\beta}$ tel que, à chaque instant d'évènement, les individus présentant un score de risque élevé $\beta^\top \mathbb{X}$ aient une probabilité plus forte d'être l'évènement observé. Autrement dit, $\widehat{\beta}$ ajuste la frontière entre "faible risque" et "fort risque" de manière à maximiser la vraisemblance des ordres d'évènements observés.

Remarque 13 (Résolution numérique). En pratique, le maximum de $\ell_p(\beta)$ est obtenu numériquement par des algorithmes de type Newton–Raphson [TG00] :

$$\beta^{(k+1)} = \beta^{(k)} + \mathcal{I}(\beta^{(k)})^{-1} U(\beta^{(k)}),$$

où $U(\beta)$ est le vecteur score (gradient de la log-vraisemblance partielle) et $\mathcal{I}(\beta)$ la matrice d'information observée (cf (8.2)). Ce schéma itératif converge vers $\widehat{\beta}$ sous des conditions classiques de régularité.

6.4 Estimation de la fonction de risque de base (méthode de Breslow)

Définition 6.4.1 (Estimateur de Breslow [Bre74])**.** *Une fois $\widehat{\beta}$ estimé, la fonction de risque cumulée de base $H_0(t)$ peut être estimée par la méthode de Breslow :*

$$\widehat{H}_0(t) = \sum_{j : t_{(j)} \le t} \frac{d_j}{\sum\limits_{i \in \mathcal{R}(t_{(j)})} \exp(\widehat{\beta}^\top \mathbb{X}_i)},$$

où d_j est le nombre d'évènements observés à l'instant $t_{(j)}$ i.e, $d_j = \sum_{i=1}^{n} \mathbb{1}_{\{T_i = t_{(j)}, D_i = 1\}}$.

Chaque terme du numérateur (d_j) correspond au nombre d'évènements réellement observés à l'instant $t_{(j)}$. Le dénominateur représente le *risque total exposé* à cet instant, pondéré par les effets estimés des covariables. La fraction

$$\frac{d_j}{\sum\limits_{i \in \mathcal{R}(t_{(j)})} \exp\Big(\widehat{\beta}^\top X_i\Big)}$$

peut être interprétée comme la contribution empirique au risque cumulatif de base.

Propriété 6.4.1 (Fonction de survie estimée)**.** *On en déduit l'estimation de la fonction de survie pour un individu i de covariables $\mathbb{X}_i$:*

$$\widehat{\mathcal{S}}(t \mid \mathbb{X}_i) = \exp\Big[-\widehat{H}_0(t)\, \exp\left(\widehat{\beta}^\top \mathbb{X}_i\right)\Big].$$

Cette expression illustre la puissance du modèle de Cox :

— la partie non paramétrique $\widehat{H}_0(t)$ capte la dynamique temporelle ;

— la partie paramétrique $\exp\left(\widehat{\beta}^\top \mathbb{X}\right)$ capte l'effet des covariables ;

— leur produit traduit la *séparabilité* du risque.

C'est cette dualité, qui lie flexibilité temporelle et linéarité covariable, qui explique la longévité et la popularité du modèle de Cox.

Chapitre 7

Modèles de Cox pénalisés

Les modèles de Cox classiques fonctionnent bien lorsque le nombre de covariables p reste modéré par rapport au nombre d'observations n. Cependant, dans de nombreux contextes modernes (données génomiques, imagerie, apprentissage automatique), on dispose de *beaucoup plus de variables explicatives que de sujets observés*. Dans ce cas, la maximisation de la vraisemblance partielle classique devient instable, voire impossible.

Une solution consiste à introduire un **terme de pénalisation** sur les coefficients β, afin de contrôler la complexité du modèle. C'est le principe des modèles de Cox pénalisés.

7.1 Principe général

Définition 7.1.1 (Log-vraisemblance partielle pénalisée)**.** *On définit la* log-vraisemblance partielle pénalisée *comme :*

$$\ell_p^{(pen)}(\beta) = \ell_p(\beta) - \lambda\, P(\beta),$$

où :

- $\ell_p(\beta)$ *est la log-vraisemblance partielle de Cox,*
- $P(\beta)$ *est un terme de pénalisation mesurant la complexité du vecteur* β*,*
- $\lambda > 0$ *est un paramètre de régularisation contrôlant l'intensité de la pénalisation.*

La pénalisation agit comme une contrainte : elle limite la taille des coefficients pour éviter le sur-apprentissage et améliorer la stabilité numérique. Le paramètre λ règle le compromis entre *biais* et *variance* :

- $\lambda = 0$: modèle de Cox classique ;
- λ grand : modèle plus lisse, coefficients plus petits, voire nuls.

7.2 Types de pénalisations usuelles

Pénalisation LASSO (ℓ_1)

Définition 7.2.1 (Cox–LASSO)**.** *La pénalisation de type LASSO (Least Absolute Shrinkage and Selection Operator), introduit par Robert Tibshirani [Tib97], consiste à ajouter la norme* ℓ_1 *des coefficients :*

$$P_{LASSO}(\beta) = \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|.$$

La log-vraisemblance partielle pénalisée devient :

$$\ell_p^{(LASSO)}(\beta) = \ell_p(\beta) - \lambda \sum_{j=1}^{p} |\beta_j|.$$

Remarque 14 (Intérêt du LASSO). Cette pénalisation encourage la **sparsité** : certains coefficients sont ramenés exactement à zéro. Elle est particulièrement utile lorsque $p \gg n$ (système dit « sous-optimal » où le nombre de coefficients de variables à estimer est bien supérieur au nombre d'observations disponibles), car elle réalise simultanément :

— une *sélection de variables* automatique ;

— une *régularisation* de la solution pour éviter le sur-ajustement.

Le LASSO est donc adapté aux données de grande dimension, où seule une petite proportion de covariables est réellement informative.

Pénalisation ridge (ℓ_2)

Définition 7.2.2 (Cox–ridge [VvH94])**.** *La pénalisation ridge repose sur la norme quadratique :*

$$P_{ridge}(\beta) = \|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2.$$

Le modèle s'écrit alors :

$$\ell_p^{(ridge)}(\beta) = \ell_p(\beta) - \lambda \sum_{j=1}^{p} \beta_j^2.$$

Remarque 15 (Intérêt du ridge). Cette pénalisation ne met pas les coefficients à zéro, mais les *réduit de manière continue*. Elle est efficace lorsque plusieurs covariables sont corrélées, car elle *stabilise les estimations* en répartissant le poids entre variables colinéaires. Le ridge privilégie donc la *robustesse* au détriment de la parcimonie.

Pénalisation élastique (Elastic Net)

Définition 7.2.3 (Cox–Elastic Net [FHT10])**.** *La pénalisation* Elastic Net *combine les effets du LASSO et du ridge :*

$$P_{EN}(\beta) = \alpha\|\beta\|_1 + \frac{(1-\alpha)}{2}\|\beta\|_2^2,$$

où $0 \leq \alpha \leq 1$ pondère la contribution respective des deux termes. La log-vraisemblance partielle pénalisée est alors :

$$\ell_p^{(EN)}(\beta) = \ell_p(\beta) - \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} \beta_j^2 \right].$$

Remarque 16 (Intérêt de l'Elastic Net). Ce compromis hérite :

— de la **sparsité** du LASSO (sélection de variables pertinentes) ;

— de la **stabilité** du ridge (gestion des covariables corrélées).

C'est le modèle de choix lorsque les variables explicatives présentent de fortes corrélations ou lorsqu'on cherche un équilibre entre sélection et robustesse.

Remarque sur le choix de λ. Le paramètre de régularisation λ est généralement choisi par *validation croisée*, en maximisant la log-vraisemblance partielle moyenne sur les échantillons de validation. Une valeur optimale de λ permet de minimiser le risque de sur-apprentissage tout en conservant les variables les plus explicatives.

Chapitre 8

Inférence et interprétation dans le modèle de Cox

On consacre cette section à l'étude de la quantification de l'incertitude associée aux coefficients estimés $\widehat{\beta}$ du modèle de Cox, ainsi que leur interprétation statistique et pratique. Le modèle de Cox, bien que semi-paramétrique, se prête à une inférence classique via la log-vraisemblance partielle et l'information observée.

8.1 Intervalles de confiance et incertitude d'estimation

Définition 8.1.1 (Estimateur et variance asymptotique)**.** *Sous les hypothèses classiques de régularité et de proportionnalité des risques, l'estimateur de maximum de vraisemblance partielle $\widehat{\beta}$ est :*

$$\widehat{\boldsymbol{\beta}} \xrightarrow{asymp.} \mathcal{N}\big(\boldsymbol{\beta}, \mathcal{I}(\widehat{\boldsymbol{\beta}})^{-1}\big), \tag{8.1}$$

où $\mathcal{I}(\widehat{\beta})$ est la matrice d'information observée*, c'est-à-dire la négative de la hessienne de la log-vraisemblance partielle :*

$$\mathcal{I}(\widehat{\beta}) = -\frac{\partial^2 \ell_p(\beta)}{\partial\beta\,\partial\beta^\top}\Big|_{\beta=\widehat{\beta}}. \tag{8.2}$$

L'écriture (8.1) implique la matrice de variance-covariance à s'exprimer comme :

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = \mathcal{I}(\widehat{\boldsymbol{\beta}})^{-1}, \quad \textit{et} \quad \widehat{\mathrm{Var}}(\widehat{\beta}_j) = \big[\mathcal{I}(\widehat{\boldsymbol{\beta}})^{-1}\big]_{jj}. \tag{8.3}$$

Cette approximation gaussienne découle du théorème de la limite centrale : lorsque la taille de l'échantillon augmente, la distribution de $\widehat{\beta}$ se concentre autour de la vraie valeur β. Ainsi, on peut quantifier l'incertitude de chaque coefficient à partir de sa variance estimée $\widehat{\mathrm{Var}}(\widehat{\beta}_j)$.

Remarque 17 (Intuition sur la matrice d'information de Fisher). La matrice d'information de Fisher mesure la *quantité d'information* que les données apportent sur les paramètres à estimer. Elle est construite à partir de la dérivée seconde (la Hessienne) de la log-vraisemblance : cette dérivée traduit la *courbure* de la fonction autour du maximum. Intuitivement :

- si la log-vraisemblance est très courbée autour de son maximum, le pic est étroit : l'estimation de $\widehat{\beta}$ est précise et la variance faible ;
- si la log-vraisemblance est plate, alors de petites variations de β produisent peu de changement dans la vraisemblance : l'incertitude sur $\widehat{\beta}$ est élevée.

Ainsi, la dérivée d'ordre deux $-\frac{\partial^2 \ell_p}{\partial \beta^2}$ quantifie *l'accélération* de la vraisemblance lorsque l'on s'éloigne du maximum. Plus cette accélération est forte, plus le modèle « réagit » aux variations des paramètres, et donc plus on dispose d'information pour les estimer avec précision.

Propriété 8.1.1 (Intervalle de confiance asymptotique)**.** *Pour chaque coefficient β_j, un intervalle de confiance asymptotique au niveau $1-\alpha$ s'écrit :*

$$\widehat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_j)},$$

où $z_{1-\alpha/2}$ est le quantile de la loi normale standard $\mathcal{N}(0,1)$ (par exemple $z_{0.975} \approx 1.96$ pour un niveau de confiance de 95%).

Preuve. Sous les hypothèses de régularité usuelles du modèle de Cox (différentiabilité, information finie, indépendance des observations et censure non-informative), l'estimateur de maximum de vraisemblance partielle $\widehat{\beta}$ vérifie l'approximation asymptotique suivante :

$$\sqrt{n}\,(\widehat{\beta}-\beta) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \mathcal{I}(\beta)^{-1}\big),$$

où $\mathcal{I}(\beta)$ désigne la matrice d'information de Fisher du modèle, définie par : $\mathcal{I}(\beta) = -\mathbb{E}\left[\frac{\partial^2 \ell_p(\beta)}{\partial\beta\,\partial\beta^\top}\right]$. Cette convergence en loi implique, pour chaque composante $j = 1, \ldots, p$:

$$\widehat{\beta}_j \approx \mathcal{N}\big(\beta_j,\, \widehat{\mathrm{Var}}(\widehat{\beta}_j)\big), \quad \text{où } \widehat{\mathrm{Var}}(\widehat{\beta}_j) = [\mathcal{I}(\widehat{\beta})^{-1}]_{jj}.$$

Dès lors, en utilisant la propriété de la loi normale standard $\mathcal{N}(0,1)$:

$$\mathbb{P}\left(-z_{1-\alpha/2} \le \frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_j)}} \le z_{1-\alpha/2}\right) \approx 1-\alpha,$$

ce qui se réécrit :

$$\mathbb{P}\left(\widehat{\beta}_j - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_j)} \le \beta_j \le \widehat{\beta}_j + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_j)}\right) \approx 1-\alpha.$$

Ainsi, l'intervalle de confiance asymptotique au niveau $1-\alpha$ pour le coefficient β_j s'écrit :

$$\widehat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_j)}.$$

□

Ces intervalles traduisent l'incertitude sur l'effet estimé d'une covariable sur le risque instantané. Plus la variance est faible, plus l'estimation de l'effet est précise. Un intervalle de confiance incluant 0 indique qu'on ne peut pas rejeter l'hypothèse d'absence d'effet significatif de la covariable sur le risque.

8.2 Significativité et interprétation des coefficients

Test de Wald

Définition 8.2.1 (Test de Wald)**.** *Le test de Wald permet de tester, pour chaque coefficient $(\beta_j)_{j\in\{1,\ldots,p\}}$, l'hypothèse :*

$$\begin{cases} H_0 : \beta_j = 0, \\ H_1 : \beta_j \neq 0. \end{cases}$$

La statistique de test est :

$$Z_j = \frac{\widehat{\beta}_j}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_j)}}, \quad \textit{avec} \quad Z_j^2 \sim \chi_1^2 \textit{ sous } H_0.$$

Le test de Wald évalue si le rapport « estimation / incertitude » est suffisamment grand pour conclure à un effet significatif. Un $|Z_j|$ élevé (ou une p-valeur faible) indique une covariable associée significativement au risque de l'évènement.

Test du rapport de vraisemblance

Définition 8.2.2 (Test du rapport de vraisemblance)**.** *On compare la log-vraisemblance partielle du modèle complet $\ell_p(\widehat{\beta})$ à celle du modèle restreint sans la covariable testée ($\ell_p(\beta_j = 0)$). La statistique de test est :*

$$\Lambda = 2\big[\ell_p(\widehat{\beta}) - \ell_p(\beta_j = 0)\big], \quad \textit{avec} \quad \Lambda \sim \chi^2_1 \textit{ sous } H_0.$$

Remarque 18 (Lien avec la vraisemblance). Ce test mesure la perte d'ajustement lorsqu'on contraint le modèle en fixant $\beta_j = 0$. Si la vraisemblance diminue fortement, cela signifie que la covariable améliore significativement la capacité prédictive du modèle.

Test de Score (Rao)

Définition 8.2.3 (Test du score (ou test de Rao))**.** *Le test du score repose sur la dérivée première de la log-vraisemblance évaluée sous H_0. La statistique s'écrit :*

$$U(\beta_0) = \frac{\partial \ell_p(\beta)}{\partial \beta}\Big|_{\beta=0}, \qquad S = U(\beta_0)^\top \mathcal{I}(\beta_0)^{-1} U(\beta_0),$$

avec $S \sim \chi^2_p$ sous H_0.

Comparaison des trois tests. Les trois tests (Wald, rapport de vraisemblance, score) sont asymptotiquement équivalents. En pratique :

- Le test de Wald est simple à interpréter mais peut être instable si les coefficients sont grands ou corrélés.
- Le test du rapport de vraisemblance est souvent plus robuste et recommandé dans les logiciels.
- Le test du score est utile lorsque le modèle complet n'a pas encore été ajusté (évaluation "locale" autour de H_0).

8.3 Interprétation des coefficients et hazard ratio

Définition 8.3.1 (Hazard ratio)**.** *Dans le modèle de Cox (6.1), le* hazard ratio *associé à une variation d'une unité de la covariable X_j s'exprime par :*

$$\mathrm{HR}_j = \exp(\beta_j).$$

Remarque 19. Une interprétation du HR :

- Si $\beta_j > 0$, alors $\mathrm{HR}_j > 1$: la covariable *augmente le risque instantané.*
- Si $\beta_j < 0$, alors $\mathrm{HR}_j < 1$: la covariable *réduit le risque instantané.*
- Si $\beta_j = 0$, alors $\mathrm{HR}_j = 1$: la covariable n'a pas d'effet sur le risque.

Ainsi, le hazard ratio mesure l'effet multiplicatif d'une covariable sur le risque à tout instant, indépendamment du temps (hypothèse de proportionnalité des risques).

Propriété 8.3.1 (Intervalle de confiance du HR)**.** *L'intervalle de confiance à $100(1-\alpha)\%$ pour HR_j s'obtient directement à partir de celui de β_j :*

$$\Big[\exp\Big(\widehat{\beta}_j - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_j)}\Big),\ \exp\Big(\widehat{\beta}_j + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_j)}\Big)\Big].$$

Cet intervalle quantifie l'incertitude sur l'effet relatif d'une covariable sur le risque. Par exemple, un $\mathrm{HR} = 1.5$ indique une augmentation de 50% du risque instantané, mais si l'intervalle de confiance inclut 1, cette augmentation n'est pas significative.

Chapitre 9

Évaluation et métriques de performance en analyse de survie

L'évaluation d'un modèle de survie constitue une étape essentielle de toute étude appliquée. Comme dans les modèles usuels et classiques de régression ou de classification, il est indispensable de mesurer la qualité d'ajustement, la capacité prédictive et la fiabilité des estimations obtenues. Cependant, la présence de données censurées rend cette évaluation plus complexe : toutes les observations ne contribuent pas de la même manière à la vraisemblance, ni aux métriques.

Ce chapitre présente les principales mesures utilisées pour quantifier la performance d'un modèle de survie, en insistant sur deux types de métriques fondamentales :

1. la **discrimination** : capacité du modèle à hiérarchiser correctement les individus selon leur risque ;
2. la **calibration** : adéquation entre les probabilités prédites et les proportions observées d'évènements.

Nous conclurons par un rappel des critères d'ajustement (AIC, BIC) et d'une comparaison avec le modèle de Kaplan–Meier.

9.1 Motivations et enjeux des métriques de performance

Définition 9.1.1 (Objectif des métriques)**.** *Une* métrique de performance *évalue la qualité prédictive d'un modèle selon un critère précis (discrimination, calibration, vraisemblance, etc.). Elle permet de comparer plusieurs modèles, d'ajuster les hyperparamètres, et d'estimer la robustesse d'un estimateur.*

Un bon modèle de survie ne doit pas seulement s'ajuster correctement aux données d'apprentissage : il doit être capable de *prédire correctement l'ordre des évènements futurs*, et de *fournir des probabilités cohérentes avec la réalité observée*. C'est précisément ce que mesurent les indices de discrimination et de calibration.

Remarque 20 (Limites et précautions). Aucune métrique n'est universellement suffisante. Un modèle peut bien discriminer les patients (C-index élevé) mais être mal calibré (Brier Score élevé). Il est donc recommandé d'évaluer simultanément ces deux aspects pour juger de la performance globale du modèle.

9.2 Discrimination

9.2.1 L'indice de concordance (*C-index*)

Définition 9.2.1 (C-index)**.** *L'indice de concordance mesure la capacité du modèle à ordonner correctement les sujets selon leur risque prédit. Formellement :*

$$\text{C-index} = \frac{\sum_{i,j} \mathbb{1}_{\{T_j<T_i\}} \cdot \mathbb{1}_{\{\widehat{\eta}_j>\widehat{\eta}_i\}} \cdot \delta_j}{\sum_{i,j} \mathbb{1}_{\{T_j<T_i\}} \cdot \delta_j},$$

où T_i et T_j sont les temps observés, δ_j l'indicateur d'évènement, et $\widehat{\eta}_i$ le score de risque estimé pour l'individu i.

Remarque 21 (Interprétation). Le C-index représente la proportion de paires d'individus pour lesquelles le modèle classe correctement le sujet à risque plus tôt comme ayant un score de risque plus élevé. Une valeur de :

— 0.5 indique une performance aléatoire ;

— 1 correspond à une discrimination parfaite.

Remarque 22 (Lien avec l'AUC). En absence de censure, le C-index se confond avec l'aire sous la courbe ROC (AUC). C'est donc une généralisation naturelle de l'AUC au cadre de la survie, adaptée aux comparaisons partielles.

9.2.2 AUC dépendante du temps

Définition 9.2.2 (AUC temporelle)**.** *Pour évaluer la discrimination à un instant donné t, on définit une AUC dépendante du temps :*

$$\widehat{\text{AUC}}(t) = \frac{\sum_{i,j} \mathbb{1}_{\{T_i>t\}} \mathbb{1}_{\{T_j\le t\}} \cdot \mathbb{1}_{\{\widehat{\eta}_j>\widehat{\eta}_i\}} \delta_j(t)}{\sum_{i,j} \mathbb{1}_{\{T_i>t\}} \mathbb{1}_{\{T_j\le t\}} \cdot \delta_j(t)}.$$

Remarque 23 (Intuition). Cette métrique évalue la capacité du modèle à distinguer, à un instant t donné, les sujets encore en vie ($T_i > t$) de ceux ayant connu l'évènement avant t. Elle fournit une mesure temporelle de la discrimination, souvent représentée sous forme de courbe $t \mapsto \widehat{\text{AUC}}(t)$.

9.3 Calibration

9.3.1 Brier Score

Définition 9.3.1 (Brier Score censuré)**.** *Le Brier Score quantifie la précision de la probabilité de survie prédite à un horizon t :*

$$\text{BS}(t) = \frac{1}{|\mathcal{V}|} \sum_{i\in\mathcal{V}} \left[\mathbb{1}_{\{T_i\le t\}} \frac{(0 - \widehat{\mathcal{S}}(t \mid \mathbb{X}_i))^2}{\widehat{G}(T_i)} \delta_i + \mathbb{1}_{\{T_i>t\}} \frac{(1 - \widehat{\mathcal{S}}(t \mid \mathbb{X}_i))^2}{\widehat{G}(t)} \right],$$

où $\widehat{G}$ désigne la fonction de survie estimée par Kaplan–Meier de la censure.

Remarque 24 (Interprétation). Le Brier Score mesure l'écart quadratique entre la survie prédite et la survie observée. Il prend des valeurs dans $[0, 1]$, et plus il est faible, meilleure est la calibration. Un modèle parfaitement calibré aurait $\text{BS}(t) = 0$ pour tout t.

9.3.2 Diagramme de calibration (Calibration Plot)

Définition 9.3.2 (Graphique de calibration)**.** *Le* calibration plot *compare, pour un horizon $t^\star$, les probabilités de survie prédites $\widehat{\mathcal{S}}(t^\star \mid \mathbb{X}_i)$ aux proportions observées d'individus effectivement en vie à cet instant. On trace typiquement :*

Probabilité prédite vs. Probabilité observée.

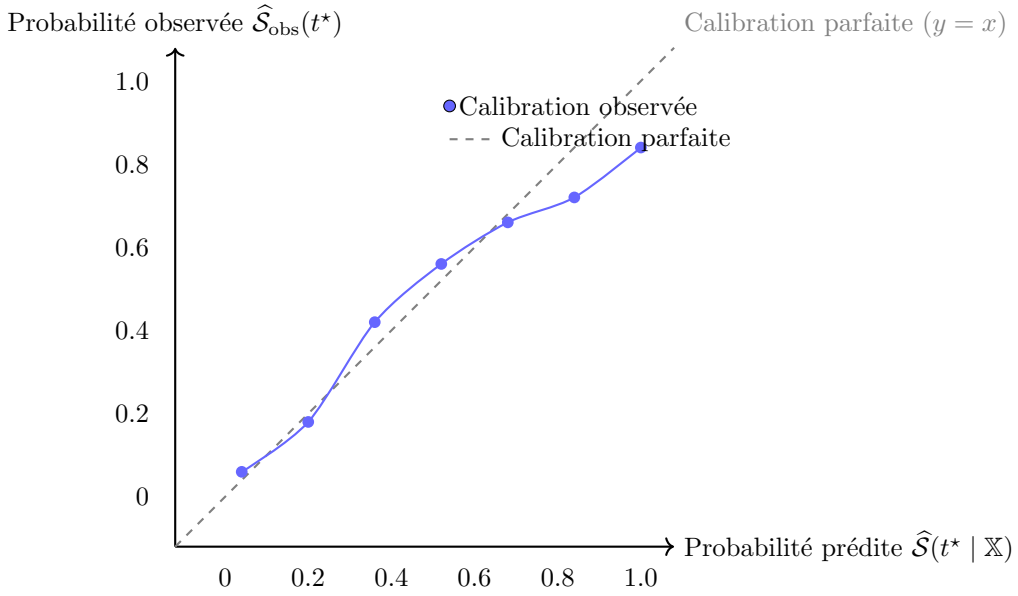

FIGURE 9.1 – Exemple de *calibration plot* : les points représentent les survies observées par groupe de patients, la ligne grise la calibration parfaite.

Remarque 25 (Lecture du graphique). Une courbe proche de la diagonale $y = x$ traduit une bonne calibration. Une courbe située au-dessus de la diagonale indique une *surévaluation du risque* (le modèle est trop pessimiste), tandis qu'une courbe en dessous indique une *sous-estimation du risque*.

9.4 Critères d'information et qualité globale d'ajustement

Définition 9.4.1 (Critères AIC et BIC)**.** *Pour comparer plusieurs modèles ajustés sur la même cohorte, on utilise souvent :*

$$\text{AIC} = -2\ell_p(\widehat{\beta}) + 2p, \qquad \text{BIC} = -2\ell_p(\widehat{\beta}) + p\log(n),$$

où p est le nombre de paramètres estimés et n le nombre d'observations.

Remarque 26 (Interprétation). Ces critères pénalisent la complexité du modèle : plus un modèle a de paramètres, plus la pénalité augmente. On choisit le modèle ayant la plus petite valeur d'AIC ou de BIC, traduisant un compromis optimal entre ajustement et parcimonie.

9.5 Comparaison modèle de Cox et Kaplan–Meier

Définition 9.5.1 (Comparaison empirique)**.** *Pour juger de la qualité du modèle de Cox, on peut comparer sa survie estimée marginale moyenne*

$$\widehat{\mathcal{S}}_{\text{Cox}}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{S}}(t \mid \mathbb{X}_i)$$

à la courbe de Kaplan–Meier non paramétrique $\widehat{\mathcal{S}}_{\text{KM}}(t)$ [HLM08].

Remarque 27 (Intérêt de la comparaison). Cette comparaison permet de vérifier si le modèle paramétrique conserve la structure empirique observée :

- une bonne concordance indique un modèle adéquat ;
- une divergence systématique suggère un mauvais ajustement ou une violation de l'hypothèse de proportionnalité des risques.

Synthèse. Les métriques de survie traduisent deux questions fondamentales :

- **Discrimination** : le modèle distingue-t-il bien les individus selon leur risque ? (C-index, AUC(t))
- **Calibration** : les probabilités prédites sont-elles cohérentes avec les fréquences observées ? (Brier Score, calibration plot)

L'évaluation d'un modèle de survie doit donc toujours articuler ces deux dimensions, complétées par une mesure d'ajustement global (AIC/BIC) et, le cas échéant, par une validation croisée sur sous-échantillons.

Chapitre 10

Application et interprétation d'un modèle de survie

Une fois un modèle de survie estimé — qu'il s'agisse d'un modèle non paramétrique de type Kaplan–Meier $\widehat{\mathcal{S}}_{\mathrm{KM}}(t)$ ou d'un modèle semi-paramétrique de Cox

$$\widehat{h}(t \mid \mathbb{X}) = \hat{h}_0(t)\, \exp(\mathbb{X}^\top \widehat{\boldsymbol{\beta}}),$$

— il devient essentiel d'interpréter, de valider et d'exploiter les résultats pour la prise de décision clinique, le suivi d'une cohorte, ou la compréhension des facteurs de risque.

10.1 Survie globale d'une cohorte

Définition 10.1.1 (Survie marginale de la cohorte)**.** *La* survie globale estimée *d'une cohorte correspond à la moyenne des probabilités de survie individuelles :*

$$\widehat{\mathcal{S}}_{\mathrm{Cox}}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{S}}(t \mid \mathbb{X}_i) = \frac{1}{n}\sum_{i=1}^{n} \exp\Big(- \int_0^t \hat{h}_0(u)\, \exp(\mathbb{X}_i^\top \widehat{\boldsymbol{\beta}})\, \mathrm{d}u \Big).$$

Remarque 28. Cette quantité est comparable à la courbe de Kaplan–Meier, qui fournit une estimation empirique non paramétrique de la survie dans la population. Comparer $\widehat{\mathcal{S}}_{\mathrm{Cox}}(t)$ à $\widehat{\mathcal{S}}_{\mathrm{KM}}(t)$ permet d'évaluer la calibration globale du modèle.

10.2 Stratification du risque

Définition 10.2.1 (Score de risque et stratification)**.** *Dans le modèle de Cox, chaque individu i se voit attribuer un* score de risque linéaire *:*

$$\eta_i = \mathbb{X}_i^\top \widehat{\boldsymbol{\beta}}.$$

Ce score traduit la tendance relative de l'individu à subir l'évènement, toutes choses égales par ailleurs. Une stratification du risque consiste à partitionner la cohorte selon des quantiles du score η_i (ex. : faible, intermédiaire, élevé).

Remarque 29. Cette approche permet de visualiser des courbes de survie distinctes par groupe de risque, et d'évaluer la capacité du modèle à discriminer entre profils pronostiques. En pratique, la stratification du risque est utilisée pour définir des stratégies thérapeutiques différenciées : intensifier le suivi ou ajuster le traitement pour les groupes à haut risque.

10.3 Interprétation des coefficients

Propriété 10.3.1 (Interprétation multiplicative)**.** *Dans le modèle de Cox, les coefficients $\widehat{\beta}_j$ s'interprètent à travers les* hazard ratios *:*

$$HR_j = \exp(\widehat{\beta}_j).$$

Un $HR_j > 1$ signifie que la variable X_j augmente le risque instantané d'évènement (facteur défavorable), tandis qu'un $HR_j < 1$ indique un effet protecteur.

Remarque 30. L'interprétation reste *conditionnelle au modèle*, c'est-à-dire à l'hypothèse de proportionnalité des risques. Les intervalles de confiance des $\widehat{\beta}_j$ et les tests de significativité (Wald, score ou rapport de vraisemblance) permettent de juger la robustesse de ces effets.

10.4 Prédiction individuelle de survie

Définition 10.4.1 (Survie individuelle prédite)**.** *Pour un individu de caractéristiques $\mathbb{X}^\star$, la fonction de survie prédite est :*

$$\widehat{\mathcal{S}}(t \mid \mathbb{X}^\star) = \exp\Big(- \int_0^t \hat{h}_0(u)\, \exp(\mathbb{X}^{\star\top} \widehat{\boldsymbol{\beta}})\, \mathrm{d}u \Big).$$

Remarque 31. En pratique, la prédiction de survie individuelle est sujette à une incertitude importante, car elle dépend à la fois de la variance des coefficients et de l'estimation non paramétrique de $\hat{h}_0(t)$. On privilégie souvent une interprétation relative (comparaison de deux profils) plutôt qu'une probabilité absolue de survie.

10.5 Extension : modèles à effets mixtes et dépendance intra-sujet

Définition 10.5.1 (Modèle de Cox à effets aléatoires)**.** *Pour des données longitudinales ou hiérarchisées (plusieurs observations par patient), on peut introduire un terme aléatoire b_i :*

$$h(t \mid \mathbb{X}_i, b_i) = h_0(t)\, \exp(\mathbb{X}_i^\top \boldsymbol{\beta} + b_i), \quad b_i \sim \mathcal{N}(0, \sigma_b^2).$$

Remarque 32. Ce type de modèle, dit *à effets mixtes*, permet de capturer l'hétérogénéité inter-individuelle et la corrélation intra-sujet. Il se situe à l'interface entre les modèles de survie et les modèles hiérarchiques bayésiens. De tels modèles nécessitent une estimation plus complexe (par EM algorithm ou MCMC), mais offrent une description plus réaliste des trajectoires patient.

Résumé du chapitre

- La courbe de Kaplan–Meier ou la survie marginale du modèle de Cox décrivent la survie globale de la cohorte.
- Le score linéaire $\eta_i = \mathbb{X}_i^\top \widehat{\boldsymbol{\beta}}$ permet une stratification du risque et une aide à la décision clinique.
- Les coefficients s'interprètent en termes de hazard ratios, mesurant l'impact relatif des covariables.
- Les prédictions individuelles sont possibles mais incertaines, surtout à long terme.
- Les extensions à effets mixtes offrent une modélisation avancée pour les données longitudinales.

Chapitre 11

Modèles à risque généralisés : au-delà du modèle linéaire

Le modèle de Cox repose sur une hypothèse fondamentale de *linéarité* du prédicteur de risque :

$$\eta(\mathbb{X}) = \mathbb{X}^\top \boldsymbol{\beta}.$$

Cette hypothèse, bien que pratique et interprétable, peut être trop restrictive lorsque la relation entre les covariables et le risque n'est pas linéaire. Il est alors naturel d'envisager des extensions où la fonction de risque dépend d'un prédicteur non linéaire :

$$\eta(\mathbb{X}) = g(\mathbb{X}),$$

tout en conservant la structure de la vraisemblance partielle.

11.1 Principe général

Définition 11.1.1 (Modèle de survie généralisé)**.** *On définit un modèle de survie à risque proportionnel généralisé par :*

$$h(t \mid \mathbb{X}) = h_0(t)\, \exp(g(\mathbb{X})),$$

où $g : \mathbb{R}^p \to \mathbb{R}$ est une fonction (éventuellement non linéaire) apprenant à estimer le score de risque *à partir des covariables.*

Remarque 33. Le rôle de $g(\mathbb{X})$ est analogue à celui du prédicteur linéaire $\mathbb{X}^\top \beta$ dans le modèle de Cox classique. La fonction g peut être paramétrique (régression logistique, spline, MLP, etc.) ou non paramétrique (forêt aléatoire, gradient boosting).

11.2 Vraisemblance partielle généralisée

Définition 11.2.1 (Vraisemblance partielle généralisée)**.** *Le principe de Cox reste valable : à chaque instant $t_{(j)}$, la probabilité que l'individu i soit celui qui subit l'évènement est donnée par :*

$$\mathbb{P}(i \text{ défaillant à } t_{(j)}) = \frac{\exp(g(\mathbb{X}_i))}{\sum_{k \in \mathcal{R}(t_{(j)})} \exp(g(\mathbb{X}_k))}.$$

La vraisemblance partielle généralisée *s'écrit donc :*

$$L_p(g) = \prod_{j=1}^{m} \frac{\exp\big(g(\mathbb{X}_{(j)})\big)}{\sum_{i \in \mathcal{R}(t_{(j)})} \exp\big(g(\mathbb{X}_i)\big)}.$$

Remarque 34. On remarque que $h_0(t)$ disparaît toujours dans le rapport : c'est cette propriété de *séparabilité du risque* qui rend le modèle de Cox robuste à la forme fonctionnelle du risque de base.

11.3 Étape d'optimisation

Définition 11.3.1 (Maximisation de la log-vraisemblance partielle)**.** *L'estimation de g repose sur la maximisation de la log-vraisemblance partielle :*

$$\ell_p(g) = \sum_{j=1}^{m} \left[g(\mathbb{X}_{(j)}) - \log\Big(\sum_{i \in \mathcal{R}(t_{(j)})} \exp(g(\mathbb{X}_i)) \Big) \right].$$

Remarque 35. L'optimisation de $\ell_p(g)$ ne nécessite pas la connaissance explicite de $h_0(t)$. Elle se prête ainsi naturellement à une optimisation numérique par gradient (pour réseaux de neurones) ou par itérations fonctionnelles (pour forêts ou boosting).

11.4 Exemples de fonctions $g(\mathbb{X})$

11.4.1 Cas linéaire classique

$$g(\mathbb{X}) = \mathbb{X}^\top \beta.$$

C'est le modèle de Cox standard. L'optimisation est convexe et donne une solution analytique via Newton–Raphson.

11.4.2 Cas forêt aléatoire

$$g(\mathbb{X}) = \log\Big(\frac{1}{T} \sum_{t=1}^{T} \exp(f_t(\mathbb{X})) \Big),$$

où chaque f_t est un arbre d'apprentissage. Ce type de modèle correspond aux *Random Survival Forests* [IKBL08], où le risque est estimé par moyennage des sous-modèles d'arbre.

11.4.3 Cas réseau de neurones

$$g(\mathbb{X}) = \mathrm{NN}_\theta(\mathbb{X}),$$

où NN_θ désigne un réseau de neurones entièrement paramétré par θ. On maximise $\ell_p(g_\theta)$ par descente de gradient, ce qui définit le modèle *DeepSurv* [KSC+18].

Remarque 36. Le terme $\mathrm{NN}_\theta(\mathbb{X})$ joue le même rôle que $\mathbb{X}^\top \beta$: il fournit un score de risque différentiable optimisé par la vraisemblance partielle.

11.4.4 Cas paramétrique

On peut spécifier une forme analytique connue, par exemple :

$$g(\mathbb{X}) = \sum_{j=1}^{p} \beta_j \sin(\omega_j X_j),$$

ou toute autre fonction paramétrique adaptée au problème physique ou biologique considéré. Ce type de modélisation conserve la flexibilité tout en gardant une interprétabilité.

11.5 Avantages et considérations pratiques

— **Flexibilité :** ces extensions permettent de capturer des relations complexes entre covariables et risque.

— **Principe invariant :** la maximisation de la vraisemblance partielle reste le cœur du processus d'estimation.

— **Comparabilité :** les scores $g(\mathbb{X})$, bien que non linéaires, conservent une interprétation monotone du risque relatif.

— **Calibration et validation :** les mêmes outils (C-index, Brier score, courbes de calibration) s'appliquent pour évaluer la performance.

Résumé du chapitre

— Le modèle de Cox peut être généralisé en remplaçant le prédicteur linéaire $\mathbb{X}^\top \beta$ par une fonction non linéaire $g(\mathbb{X})$.

— La clé reste la maximisation de la log-vraisemblance partielle, indépendante du choix du modèle pour g.

— Ces extensions permettent d'explorer des formes de dépendance plus riches et d'améliorer la performance prédictive.

Chapitre 12

Implémentation pratique des modèles de survie

L'analyse de survie, bien qu'ancrée dans la théorie statistique, dispose aujourd'hui d'outils logiciels matures et fiables pour l'estimation des courbes de Kaplan–Meier, des modèles de Cox, et de leurs versions pénalisées. Cette section présente les principaux environnements utilisés en pratique : Python, R et SAS.

12.1 Implémentation en Python

Le langage **Python** propose plusieurs bibliothèques spécialisées dans l'analyse de survie, adaptées aussi bien à la recherche qu'à la production.

Package `lifelines`

- `KaplanMeierFitter` — estimation et tracé de la courbe de survie de Kaplan–Meier.
- `CoxPHFitter` — estimation du modèle de Cox (vraisemblance partielle classique).
- `AalenAdditiveFitter` — modèle additif d'Aalen.
- `plot_survival_function` — visualisation de $\widehat{\mathcal{S}}(t)$.

Package `sksurv` (scikit-survival)

- `KaplanMeierEstimator` — estimateur non paramétrique.
- `CoxPHSurvivalAnalysis` — modèle de Cox sous API compatible `scikit-learn`.
- `BreslowEstimator` — estimation du risque cumulatif de base.
- `concordance_index_censored` — mesure de performance (C-index).

Package `skglm`

- `GeneralizedLinearEstimator` — API générique pour les modèles linéaires régularisés.
- `CoxLasso` et `CoxElasticNet` — modèles de Cox pénalisés par LASSO ou Elastic Net.
- `path_enet` — calcul du chemin de régularisation complet.

Remarque 37. En pratique :

— `lifelines` est idéal pour la visualisation et les analyses simples.

— `sksurv` s'intègre parfaitement avec `scikit-learn` pour les pipelines prédictifs.

— `skglm` est le plus performant pour les modèles pénalisés (implémentation optimisée en C++).

12.2 Implémentation en R

Le langage **R** demeure la référence historique et académique pour l'analyse de survie, avec un écosystème riche et stable.

Package `survival`

- `survfit()` — estimation de Kaplan–Meier.
- `coxph()` — estimation du modèle de Cox (classique et stratifié).
- `survreg()` — modèles paramétriques (Weibull, exponentiel...).
- `basehaz()` — estimation du risque de base (méthode de Breslow).

Package `glmnet`

- `glmnet()` — estimation des modèles linéaires pénalisés (LASSO, ridge, Elastic Net).
- `cv.glmnet()` — sélection automatique du paramètre de régularisation λ par validation croisée.
- Compatible avec la famille `"cox"` pour la régression de Cox pénalisée.

Package `survminer`

- `ggsurvplot()` — visualisation améliorée des courbes de survie (basée sur `ggplot2`).
- `surv_compare()` — comparaison de groupes (test du log-rank).

12.3 Implémentation en SAS

Le logiciel **SAS** offre une implémentation robuste et optimisée de l'analyse de survie, particulièrement utilisée dans les milieux biomédicaux et pharmaceutiques.

Procédures principales

- `PROC LIFETEST` — estimation de Kaplan–Meier, tests du log-rank, graphiques de survie.
- `PROC PHREG` — modèle de Cox (avec options pour stratification, effets aléatoires, et pénalisation).
- `BASELINE` — estimation du risque de base et des courbes de survie ajustées.

SAS intègre des options avancées de gestion des censures, des poids et des covariables dépendantes du temps, avec une documentation exhaustive.

12.4 Synthèse d'emploi des packages

Langage	Packages principaux	Fonctions clés	Références
Python	`lifelines`, `sksurv`, `skglm`	Estimation de Kaplan–Meier, modèle de Cox, variantes pénalisées (LASSO, Elastic Net)	[DPKJea19], [Pö20], [MGS23]
R	`survival`, `glmnet`, `survminer`	Kaplan–Meier, modèle de Cox, Cox pénalisé, visualisation et comparaison de groupes	[TG00], [FHT10], [KKB17]
SAS	`PROC LIFETEST`, `PROC PHREG`	Estimation de Kaplan–Meier, modèle de Cox, estimation de Breslow	[All10]

Table 12.1 – Principaux environnements logiciels pour l'analyse de survie

Chapitre 13

Travaux Pratiques : Application de l'analyse de survie avec Python

13.1 Objectifs du TP

Ce TP a pour but de mettre en pratique les méthodes vues dans le cours à partir de données réelles. L'étudiant devra :

- explorer un jeu de données censuré (variables explicatives, durée, statut) ;
- estimer et interpréter une fonction de survie par la méthode de Kaplan–Meier ;
- ajuster un modèle de Cox proportionnel et interpréter ses coefficients ;
- évaluer la qualité du modèle par des métriques de discrimination et de calibration ;
- comparer empiriquement les résultats entre l'approche non paramétrique et le modèle semi-paramétrique.

13.2 Données utilisées

Définition 13.2.1 (Jeu de données proposé)**.** *On utilisera ici le jeu de données* Lung Cancer Dataset*, inclus dans la librairie* `lifelines`*. Chaque observation correspond à un patient atteint d'un cancer du poumon, avec :*

- *la durée de suivi (*`time`*) ;*
- *l'évènement observé (*`status`*) : 1 = décès, 0 = censure ;*
- *des covariables : âge, sexe, score de performance, etc.*

Chargement des données :

```
from lifelines.datasets import load_lung
import pandas as pd

df = load_lung()
df.head()
```

13.3 Partie I : Exploration et analyse descriptive

(i) Décrire le jeu de données : nombre d'observations, proportion de censures, statistiques descriptives.

(ii) Étudier la distribution de la variable de durée (`time`) et du statut (`status`).

(iii) Identifier les variables susceptibles d'influencer la survie.

(iv) Formuler des hypothèses qualitatives sur les facteurs de risque.

Objectif pédagogique : comprendre le rôle des covariables avant toute modélisation.

13.4 Partie II : Estimation de Kaplan–Meier

(i) Estimer la fonction de survie globale :

```
from lifelines import KaplanMeierFitter
import matplotlib.pyplot as plt

kmf = KaplanMeierFitter()
kmf.fit(df["time"], event_observed=df["status"])
kmf.plot_survival_function(ci_show=True)
plt.title("Courbe de Kaplan{Meier avec intervalle de confiance")
plt.xlabel("Temps (jours)")
plt.ylabel("Probabilité de survie estimée")
plt.show()
```

(ii) Diviser la population selon une variable (ex. sexe) et tracer les courbes :

```
for sex, group in df.groupby("sex"):
    kmf.fit(group["time"], event_observed=group["status"], label=f"Sexe {sex}")
    kmf.plot_survival_function(ci_show=True)
plt.title("Comparaison des courbes de Kaplan{Meier selon le sexe")
plt.show()
```

(iii) Appliquer le **test du log-rank** :

```
from lifelines.statistics import logrank_test

results = logrank_test(
    df.loc[df.sex==1, "time"], df.loc[df.sex==2, "time"],
    event_observed_A=df.loc[df.sex==1, "status"],
    event_observed_B=df.loc[df.sex==2, "status"]
)
results.print_summary()
```

(iv) Discuter des hypothèses : indépendance, censure non-informative, proportionnalité des risques.

13.5 Partie III : Modèle de Cox proportionnel

(i) Séparer les données en ensembles d'entraînement et de test :

```
from sklearn.model_selection import train_test_split

train, test = train_test_split(df, test_size=0.3, random_state=42)
```

(ii) Ajuster le modèle de Cox :

```
from lifelines import CoxPHFitter

cph = CoxPHFitter()
cph.fit(train, duration_col="time", event_col="status")
cph.print_summary()
```

(iii) Identifier les variables significatives, interpréter les coefficients et leurs intervalles de confiance.

(iv) Évaluer la performance du modèle :

```
from lifelines.utils import concordance_index

c_index = concordance_index(
    test["time"],
    -cph.predict_partial_hazard(test),
    test["status"]
)
print(f"C-index = {c_index:.3f}")
```

(v) Tracer la survie moyenne estimée :

```
surv_pred = cph.predict_survival_function(test)
mean_surv = surv_pred.mean(axis=1)
plt.plot(mean_surv.index, mean_surv.values, label="Cox marginalisé", color="blue")
plt.title("Survie marginalisée du modèle de Cox")
plt.xlabel("Temps")
plt.ylabel("Probabilité de survie moyenne")
plt.legend()
plt.show()
```

13.6 Partie IV : Comparaison avec Kaplan–Meier

(i) Comparer graphiquement :

$$\widehat{\mathcal{S}}_{\text{Cox}}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{S}}(t \mid \mathbb{X}_i)$$

à la courbe $\widehat{\mathcal{S}}_{\text{KM}}(t)$:

```
kmf.fit(test["time"], event_observed=test["status"], label="Kaplan{Meier")
kmf.plot(ci_show=True)
plt.plot(mean_surv.index, mean_surv.values, label="Cox marginalisé", color="blue")
plt.title("Comparaison Kaplan{Meier vs Cox PH")
plt.legend()
plt.show()
```

(ii) Calculer et commenter le Brier score et la calibration du modèle :

```
from sksurv.metrics import brier_score
# (Exemple si le jeu de données est adapté)
```

13.7 Pour aller plus loin (optionnel)

```
from skglm import CoxPHSurvivalAnalysis

model = CoxPHSurvivalAnalysis(alpha=0.1, l1_ratio=1.0)  # LASSO
model.fit(train.drop(columns=["time", "status"]),
          (train["status"], train["time"]))
```

Comparer les performances avec le modèle non régularisé.

Conclusion générale

L'analyse de survie occupe une place essentielle dans la modélisation statistique des phénomènes temporels où l'on s'intéresse au temps jusqu'à la survenue d'un évènement d'intérêt. Tout au long de ce cours, nous avons vu que sa spécificité réside dans sa capacité à exploiter *toute* l'information disponible — y compris celle des individus dont l'évènement n'a pas encore été observé, grâce au traitement rigoureux de la **censure**.

Les méthodes classiques de régression ou de classification, en revanche, supposent une observation complète de la variable cible, et échouent à gérer la nature partielle et temporelle de ces données. L'analyse de survie, par les estimateurs de **Kaplan–Meier** et les modèles de **Cox à risques proportionnels**, offre un cadre unifié et cohérent pour quantifier les durées, les risques, et leurs dépendances aux covariables.

Même en l'absence de censure, ces outils conservent tout leur intérêt : ils permettent d'étudier la distribution temporelle des évènements, d'interpréter les effets multiplicatifs des covariables sur le risque, et de comparer des groupes de manière dynamique. Ainsi, leur usage dépasse le cadre médical ou industriel classique : ils s'appliquent à toute situation où la *dynamique temporelle du risque* importe plus que la simple valeur moyenne d'un temps d'occurrence.

Sur le plan méthodologique, le modèle de Cox illustre la puissance des approches **semi-paramétriques**, conciliant flexibilité et interprétabilité. Ses extensions pénalisées (LASSO, Ridge, Elastic Net) et ses variantes modernes (forêts aléatoires, réseaux neuronaux, modèles mixtes) montrent que l'analyse de survie est aujourd'hui au cœur des développements statistiques et du *machine learning* appliqué aux données temporelles.

L'étudiant ayant suivi ce cours doit désormais être capable :

- d'identifier les contextes où la censure intervient et où l'analyse de survie s'impose ;
- d'appliquer et d'interpréter une estimation de Kaplan–Meier et un test du log-rank ;
- de construire et d'analyser un modèle de Cox, en comprenant la signification des coefficients, la logique de la vraisemblance partielle et les métriques de qualité associées (C-index, Brier score, calibration) ;
- d'interpréter les résultats de manière critique, en les reliant à une prise de décision ou à une stratégie de risque.

L'analyse de survie ne se limite donc pas à prédire une durée : elle permet de comprendre la structure du risque dans le temps, d'exploiter l'information incomplète, et d'en extraire une connaissance statistique utile à la décision. Elle illustre parfaitement la philosophie de la statistique moderne : *modéliser l'incertitude pour mieux éclairer l'action.*

Bibliographie

[All10] Paul D. Allison. *Survival Analysis Using SAS : A Practical Guide*. SAS Institute Inc., Cary, NC, 2nd edition, 2010.

[Bre74] Norman E. Breslow. Covariance analysis of censored survival data. *Biometrika*, 61(3) :579–594, 1974.

[Cox72] David Roxbee Cox. Regression models and life-tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, 34(2) :187–202, 1972.

[Cox75] David Roxbee Cox. Partial likelihood. *Biometrika*, 62(2) :269–276, 1975.

[DPKJea19] Cameron Davidson-Pilon, Jonas Kalderstam, Paul Jacobson, and et al. lifelines : survival analysis in python. *Journal of Open Source Software*, 4(40) :1317, 2019.

[FHT10] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1) :1–22, 2010.

[Gre26] Major Greenwood. *The Natural Duration of Cancer*. Number 33 in Reports on Public Health and Medical Subjects. His Majesty's Stationery Office (H.M.S.O.), London, 1926.

[HLM08] David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis : Regression Modeling of Time-to-Event Data*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2008.

[IKBL08] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests for r. *Bioinformatics*, 24(11) :1363–1371, 2008.

[KKB17] Alboukadel Kassambara, Marcin Kosinski, and Przemyslaw Biecek. survminer : Drawing survival curves using 'ggplot2'. *R package version 0.4.4*, 2017.

[KM58] Edward L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282) :457–481, 1958.

[KSC+18] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv : personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18 :24, 2018.

[Man66] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50 :163–170, 1966.

[MGS23] Mathurin Massias, Alexandre Gramfort, and Joseph Salmon. skglm : Efficient generalized linear models with regularization. *Journal of Machine Learning Research*, 24(334) :1–8, 2023.

[PP72] Richard Peto and Julian Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society : Series A (General)*, 135(2) :185–207, 1972.

[Pö20] Sebastian Pölsterl. scikit-survival : A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212) :1–6, 2020.

[TG00] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data : Extending the Cox Model*. Statistics for Biology and Health. Springer, 2000.

[Tib97] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4) :385–395, 1997.

[VvH94] Pierre J. M. Verweij and Hans C. van Houwelingen. Penalized likelihood in cox regression. *Statistics in Medicine*, 13(23–24) :2427–2436, 1994.