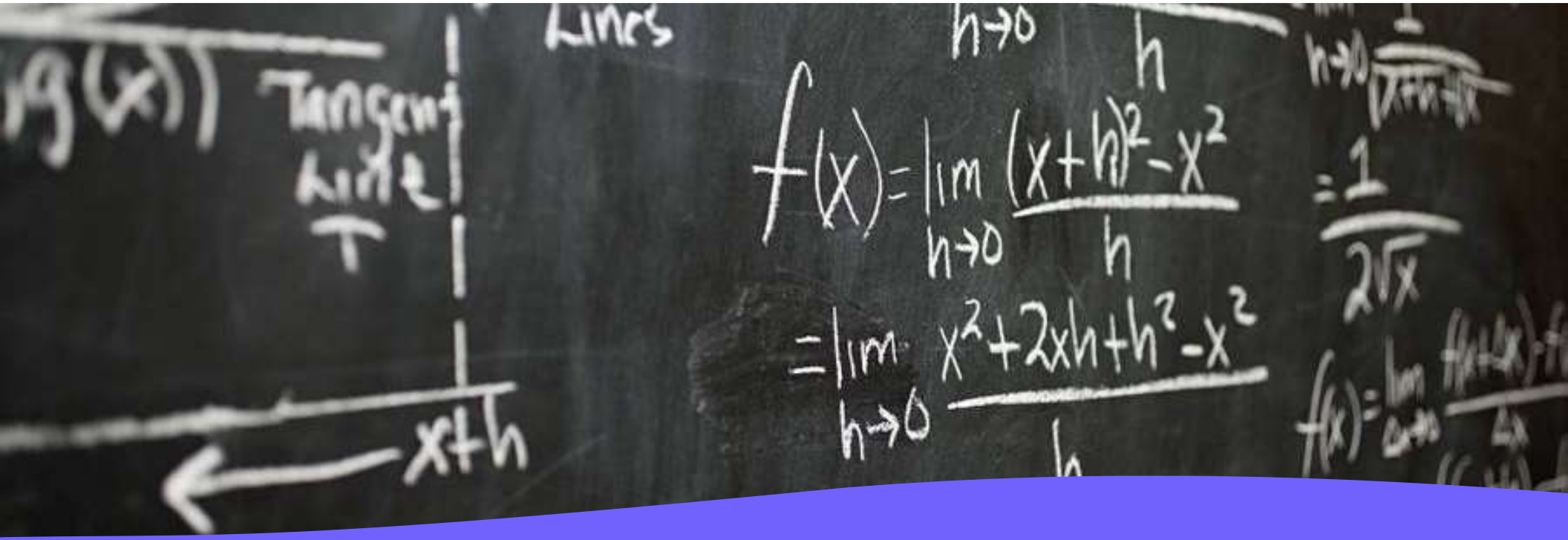


Application de l'analyse de survie à la modélisation du risque

Formation - Laboratoire d'Ingénierie Financière de L'université Laval (LABIFUL),

Présenté par: Josiane Guedem ISE, MBA

21 février 2025

Plan de la présentation

I. Notions fondamentales

1. Les données de survie.
2. La censure
3. Fonctions de densité, de décès/survie et de risque.

II. Application à la gestion du risque

1. Contexte.
2. Approche non paramétrique
3. Approche semi-paramétrique
4. Approche paramétrique

III. Cas pratique

1. Introduction
2. Structure des données et mesure du temps de survie
3. Formulation du modèle et sélection des variables
4. Résultats de la modélisation sous SAS

Introduction générale

- En analyse de survie, on utilise des modèles de survie pour étudier le temps écoulé avant qu'un événement ne survienne.
- Utilisée pour la première fois au cours des années 50, dans le domaine médical.
- Historiquement, le principal événement est le décès de l'individu, c'est pourquoi on parle généralement de survie et de décès.
- Au fil des années, l'utilisation du modèle en recherche médicale s'est étendue à d'autres situations, et l'événement peut donc être de quelconque nature : récidive d'une maladie ou à l'inverse, d'une guérison.
- L'analyse de survie aujourd'hui s'applique principalement non seulement en recherche médicale, mais aussi en sciences humaines et sociales.
- L'analyse de survie a connu de l'essor en gestion des risques avec les normes IFRS9 qui recommandent de calculer les probabilités de défaut sur toute la durée de vie des prêts (*lifetime*) et non plus sur un horizon limité d'1 an.

Objectif de la formation

- Introduire l'utilisation de l'analyse de survie dans le domaine de la gestion des risques, puis illustrer dans un cas pratique.

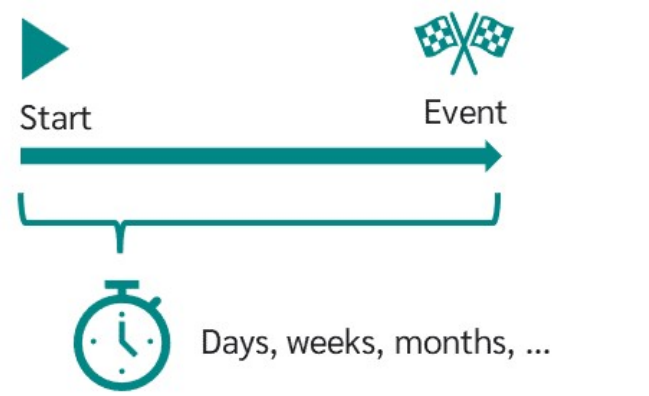
Analyse de survie: Notions fondamentales

1. Les données de survie
2. La censure
3. Fonctions de densité, de décès/survie et de risque.

1. Les données de survie

- Contrairement aux approches traditionnelles de modélisation, la variable T à modéliser en analyse de survie correspond au **temps de survie**, c'est à dire la durée d'un processus ou le temps écoulé avant la survenue d'un évènement;
- C'est pourquoi on parle aussi de **modèle de durée**.
- Principales applications en recherche médicale, en sciences humaines et sociales, par exemple:

1. La durée de vie des patients atteints d'un cancer;
2. Le temps écoulé avant la récidive d'un toxicomane;
3. L'âge d'entrée dans la vie active pour les jeunes de la génération Z;
4. Temps écoulé avant la défaillance du moteur des véhicules neufs;
5. La durée du mariage;
6. Temps écoulé avant la faillite d'une entreprise, etc.



1. Les données de survie (2)

Principales caractéristiques des variables de survie

1. Le temps de survie T est une variable continue qui ne peut prendre que des valeurs positives:

- Par conséquent, sa distribution présente généralement une forte asymétrie positive, et s'écarte ainsi de la loi normale.
- En analyse de survie, l'hypothèse de loi normale, sous-jacente a la plupart des méthodes traditionnelles de modélisation des variables continues, n'est donc plus valide;

2. T n'est pas forcément observé/mesuré pendant tout le temps que dure l'étude de l'évènement:

- On dit alors que la variable T est **"censurée"** (ou partiellement connue, c'est-à-dire que les l'information est incomplète).

2. La Censure

- La censure est un élément fondamental en analyse de survie, qui ne pourrait pas être convenablement géré avec les méthodes classiques.
- Pour des raisons de ressources (temps, finances, etc.), l'étude du phénomène ne peut pas durer indéfiniment, mais s'étend sur une période limitée.
- Chaque étude a donc une date de début et une date de fin.
- Entre lesdites date de début et de fin, l'évènement n'est pas toujours observé avec exactitude, par exemple si certains individus sont perdus de vue au cours de la période.
- On parle de **"censure",** ou **"information incomplète " .**
- Les méthodes classiques de modélisation sont inadaptées en cas de censure.

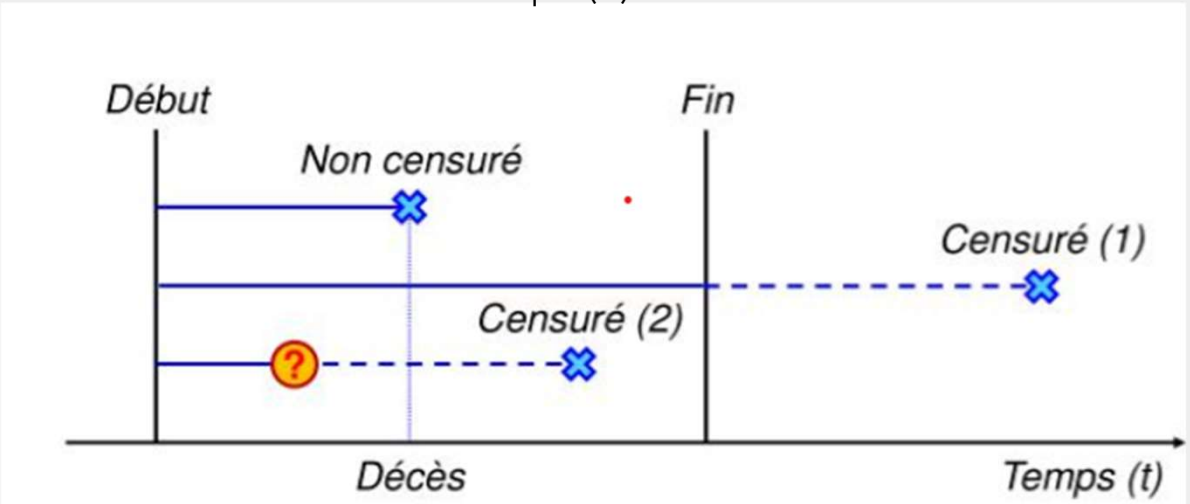
Médecine: étude du temps de survie des patients hospitalisés à cause d'une pathologie (cancer, avc, etc…)

- Le temps de survie des patients est le temps écoulé entre l'hospitalisation et le décès
- Ce temps de survie ne sera pas totalement connu :
  - Si le patient **décède après** la date de fin de l'étude (1)
  - **Son suivi est interrompu** pour plusieurs raisons indépendantes de l'étude (2)

2. La Censure (2)

Gestion des risques: étude de la survie avant le défaut

- Le défaut (manquement aux obligations financières envers la Banque) n'intervient pas obligatoirement pour tous les clients au cours de l'étude.
- Le temps de survie ne sera pas connu avec exactitude si le client:
  - Fait défaut après la fin de la période d'analyse (1)
  - Est perdu de vue sans faire défaut, par exemple s'il ferme son compte avec la Banque (2)



3. Fonctions de densité, de décès/survie et de risque

T est la variable aléatoire (continue et positive) qui définit le temps de survie avant la survenue d'un évènement:

- **La fonction de densité f(t)** décrit la probabilité instantanée que l'évènement survienne pendant un laps de temps Δt

$$f(t) = \lim_{\Delta t \to 0^+} \frac{P(t < T \leq t + \Delta t)}{\Delta t} = \frac{F(t + \Delta t) - F(t)}{dt} = \frac{dF(t)}{dt}$$

En gestion de risque, f(t) correspondra au taux de défaut entre t et t+Δt, mesuré par nombre de défauts survenus pendant le laps de temps Δt, divisé par le nombre de clients du portefeuille a l'origine t.

3. Fonctions de densité, de décès/survie et de risque (2)

- La **fonction de répartition F(t)** décrit de La probabilité d'observer un temps de survie inférieur ou égal à t

$$F(t) = Pr(T \leq t)$$

On parle également de **fonction de décès**

Si on intègre la fonction de densité f(t) en continue entre 0 et t, on obtient F(t), c a d

$$F(t) = \int_0^t f(t)dt$$

On peut par exemple utiliser la fonction de décès pour déterminer la probabilité de survivre jusqu'à 365 jours (ou 1 an). Ceci correspondra en gestion de risque a la probabilité de faire défaut sur un horizon de 1 an.

3. Fonctions de densité, de décès/survie et de risque (3)

- **La fonction de survie S (t)** décrit, contrairement à la fonction de décès, la probabilité de survivre au-delà du temps t, c a d

$$S(t) = Pr(T>t) = 1 - F(t)$$

- **La fonction de risque h(t)** décrit la probabilité instantanée que l'événement se produise à l'instant t, sachant que l'individu a survécu jusqu'à cet instant t.

$$h(t) = \lim_{\Delta t \to 0^+} \frac{Pr(t < T \leq t + \Delta t | T > t)}{\Delta t}$$

Généralement, l'objectif principal de l'analyse de survie est de modéliser le temps de survie T ou la fonction de risque h(t), qui est liée à la fonction de décès et la fonction de survie par la relation suivante :

$$h(t) = \frac{f(t)}{S(t)}$$

En gestion du risque, la fonction de risque représente le taux de défaut instantané entre t et t+Δt sachant que le temps de survie *T* est supérieur à t.

Application a la gestion du risque

1. Contexte
2. Approche non paramétrique
3. Approche semi-paramétrique
4. Approche paramétrique

1. Contexte

- En gestion des risques, on veut évaluer l'impact de divers facteurs de risque X_1, …, X_k sur le temps de survie T des clients avant le défaut.
  - **Client particulier:** variables de bureau de crédit, utilisation du crédit, habitudes de remboursement, revenu, statut professionnel, etc
  - **Client commercial:** ratios financiers (rentabilité, endettement, liquidité) et informations qualitatives (qualité du management, concentration des produits ou de la clientèle…)
- **Question:** quels sont les facteurs spécifiques au client, ou les facteurs systématiques de la macroéconomie qui peuvent influencer le temps de survie du client?
- **Hypothèse :** si les conditions macroéconomiques sont favorables et que le client a un profil peu risqué, sa capacité financière sera élevée et son pronostic de survie avant le défaut sera favorable (c'est-à-dire sa probabilité de défaut va diminuer).
- La forme fonctionnelle de la relation causale entre T et les variables explicatives dépend de la loi de probabilité de T peut s'écrire comme sui (qui peut être connue ou non *a priori*):

$$Log(T) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$ (**exemple du modèle exponentiel**)

- Plus β_k est élevé, plus le facteur influence la survie du client.
- Le modèle permettra ensuite de prédire le temps de survie pour de nouveaux clients.

2. Approche non paramétrique

- Permet une exploration rapide et simple des données de survie à partir de l'observation empirique, sans formuler a priori d'hypothèse sur la forme analytique de la distribution de probabilité de la variable de survie T.
- Très utilisé en médicine pour décrire la survie des patients car aucune une loi théorique ne permet de décrire le temps de décès de la vie humaine.
- La fonction de survie S(t) est simplement estimée de façon empirique à partir des données, en utilisant la formule dite de Kaplan Meier.

S(t) est estimée par : $S(t_i) = \prod_{j=1}^{i} Pr(T > t_j | T \geq t_j) = \prod_{j=1}^{i} \left(1 - \frac{d_j}{n_j}\right)$, avec

d_j=nombre de défauts survenus à l'instant t_j

n_j=nombre de clients qui n'étaient pas en défaut (c'est-à-dire "en vie ") juste avant la date t_j

👉**s'applique mal à l'analyse prédictive dont le but est de déterminer la relation causale entre T un et ensemble de variables explicatives.**

2. Approche non paramétrique (2)

Étude du temps de survie avant le défaut

t_j	n_j	d_j	q_j	$\widehat{S}(t_j)$
0	10	0	0	$10/10 = \mathbf{1}$
3	10	1	0	$1 * (10-1)/10 = \mathbf{0.9}$
6	9	2	1	$0.9 * (9-2)/9 = \mathbf{0.7}$
9	6	1	0	$0.7 * (6-1)/6 = \mathbf{0.58}$
10	5	1	1	$0.58 * (5-1)/5 = \mathbf{0.47}$
16	3	1	0	$0.47 * (3-1)/3 = \mathbf{0.31}$
17	2	1	0	$0.31 * (2-1)/2 = \mathbf{0.16}$
18	1	1	0	$0.16 * (1-1)/1 = \mathbf{0}$

Proportion de survivants à t_j

q_j : *nombre d'observations censurées entre* t_j *et* t_{j+1}

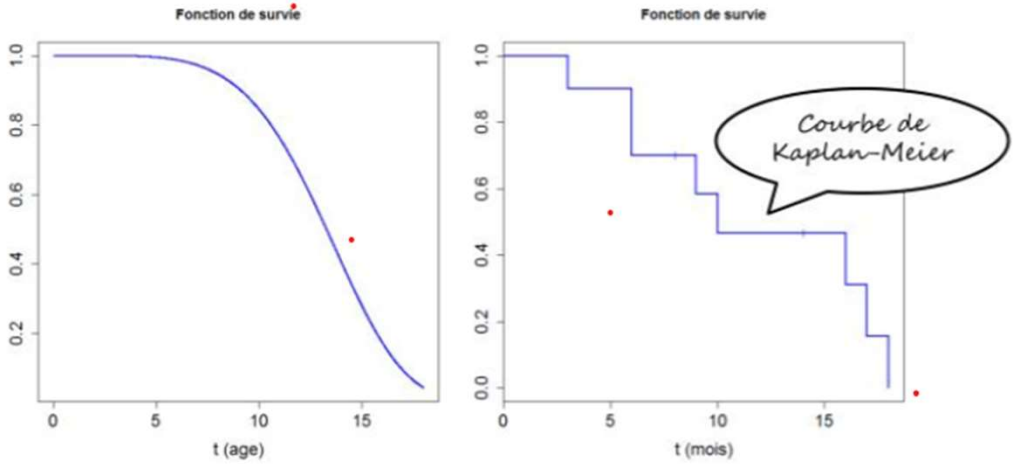
La fonction de survie empirique peut encore s'écrire: $S(t_i) = \left(1 - \frac{d_i}{n_i}\right) \times \prod_{j=1}^{i-1}\left(1 - \frac{d_j}{n_j}\right)$

$= \underbrace{\left(1 - \frac{d_i}{n_i}\right)}_{Survivants\ en\ t_i} \times S(t_{i-1})$

Fonction de survie

$S(t) = P(T > t)$

Théorie : $S(t)$ *Pratique* : $\hat{S}(t)$



3. Approche semi-paramétrique

- En estimation semi paramétrique, le **modèle de Cox** est une approche largement répandue, en particulier en médecine et sciences humaines pour modéliser le risque de décès h(t) des individus en fonction d'une combinaison de facteurs X_1, …, X_k
- La fonction de risque est le produit de deux composantes: $h(t) = \underbrace{h_0(t)}_{Risque\ de\ base} \times \underbrace{e^{\beta_1 X_{k1} + \cdots + \beta_1 X_k}}_{terme\ a\ modéliser}$
- L'approche est donc dite **" semi-paramétrique"** car :
  - Il n'est pas nécessaire de formuler une hypothèse sur la loi de $h_0(t)$ qui dépend uniquement du temps.
  - Le terme à modéliser, totalement indépendante du temps, a une forme exponentielle.
- De plus, l'âge et le sexe doivent automatiquement être ajoutés parmi les variables explicatives, car incontournables pour expliquer le risque de décès des individus

Contraintes sur la forme analytique de la fonction de risque et sur les variables explicatives. s'applique mal dans un cadre ou la loi exponentielle n'est pas en adéquation avec les données, et en l'absence de relation causale entre le risque et le sexe et/ou l'âge.

4. Approche paramétrique

Objectif:

- Résumer h(t) et donc S (t) par une fonction mathématique connue.
- Et un ou plusieurs paramètres simples qui suffisent à résumer cette fonction.

En pratique:

- Très peu adapté pour l'humain.
- Très utilisé en gestion des risques pour modéliser le temps de survie avant le défaut.

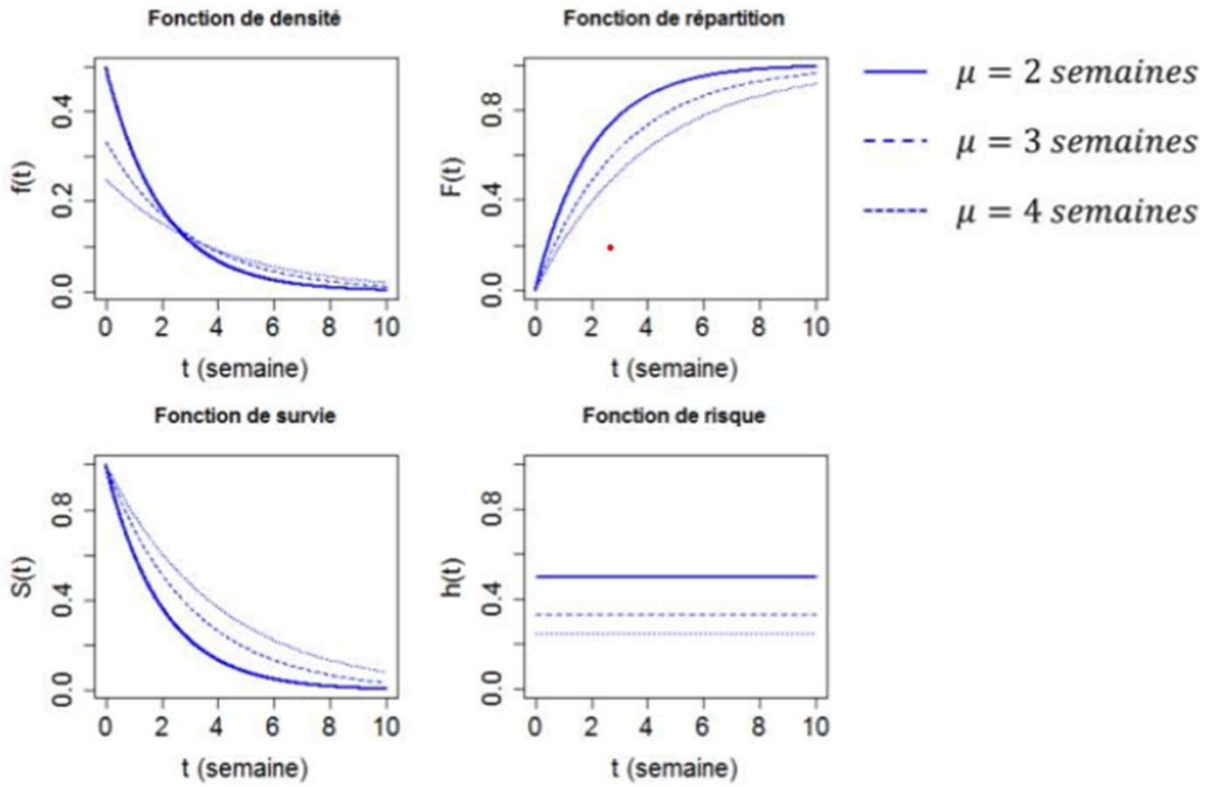
Exemple: loi exponentielle,

- Modèle paramétrique le plus simple.
- Soit μ la durée moyenne de survie d'un individu selon cette loi.

Fonction	Notation	Formulation
Densité	f(t)	$\frac{1}{\mu} e^{-\left(\frac{1}{\mu}t\right)}$
Décès	F(t)	$1 - e^{-\left(\frac{1}{\mu}t\right)}$
Survie	S(t)	$e^{-\left(\frac{1}{\mu}t\right)}$
Risque	h(t)	$\frac{1}{\mu}$

4. Approche paramétrique (2)

Loi exponentielle: Fonction de densité, décès de survie et de risque



Autres exemples de lois paramétriques:

- Weibull
- Gamma
- Log-normale
- Log-logistique

Probabilité de défaut 1an: $Pr(T \leq 52)$

car 1 an=52 semaines

Cas pratique

1. Introduction
2. Structure des données et mesure du temps de survie
3. Formulation du modèle et sélection des variables
4. Résultats de la modélisation sous SAS

1. Introduction

- Nous souhaitons utiliser l'analyse de survie pour évaluer les facteurs qui influencent la probabilité que les clients fassent défaut sur leur carte de crédit
- Pour ce faire, nous disposons d'un échantillon de 100 comptes de carte de crédit (**données fictives**) observés sur la période entre le 1er janvier 2018 et le 31 décembre 2023
- Les variables considérées comme susceptibles d'expliquer le défaut des clients (ou variables candidates) sont:
    1. L'âge du détenteur de la carte de crédit;
    2. Son statut d'emploi (chômeur ou non);
    3. Son revenu annuel, et
    4. Le nombre de jours de délinquance observé sur la carte de crédit.
- Il y a défaut lorsque le client accumule plus de 90 jours de retard de paiement du solde de sa carte (90+ jours de délinquance).

Hypothèse générale: la capacité financière du client sera d'autant plus élevée et son pronostic de survie avant le défaut d'autant plus favorable (c'est-à-dire une diminution de la probabilité de faire défaut sur sa carte de crédit) que son profil est peu risqué.

1. Introduction (2)

- **Hypothèses spécifiques sur la relation pressentie entre les variables candidates et le risque (probabilité de défaut ou pronostic de survie)**

Variable	Définition	Signe attendu (par rapport à la survie avant le défaut) [1]	Intuition économique
Age	Âge du détenteur de la carte de crédit	+	La propension à dépenser diminuerait avec l'âge, d'où une faible accumulation de dette et donc une tendance à l'augmentation du temps de survie avant le défaut.
Chômage	Statut à l'emploi du détenteur de la carte de crédit (1 s'il est au chômage et 0 sinon)	-	Le client au chômage aura plus tendance à faire défaut sur ses paiements de carte de crédit (diminution du temps de survie avant le défaut) que le client en emploi, compte tenu de ses ressources financières limitées.
Jrs délinquance	Nombre de jours de retard de paiement du solde de la carte de crédit	-	Plus les jours de retard de paiement du solde s'accumuleront, plus le défaut sera probable et par conséquent le temps de survie du client avant le défaut sera faible.
Revenu	Revenu annuel du détenteur de la carte de crédit (en CAD)	+	Des revenus plus élevés indiquent une grande capacité financière et donc une augmentation du temps de survie avant le défaut.

[1] Un signe positif implique que le temps de survie avant le défaut augmente lorsque la variable augmente (diminution de la probabilité de défaut).

☛ **Nous nous servirons des données disponibles pour confirmer ou infirmer ces hypothèses**

2. Structure des données et mesure du temps de survie

La mise en œuvre de l'analyse de survie nécessite que les éléments suivants soient connus ou disponibles dans l'échantillon de données utilisé pour la modélisation:

1. **La période d'analyse (observation des comptes de l'échantillon):**
   - Elle s'étend du début (1er janvier 2018) à la date de fin de l'analyse (31 décembre 2023)
   - La date de fin de l'analyse permet de calculer le temps de survie pour des comptes dont le défaut n'a pas été observé au cours de la période d'observation
2. **La date d'observation de chaque compte:**
   - Date à laquelle le compte commence à être considéré dans l'analyse de survie
3. **La date de fermeture de chaque compte**
   - Pour les comptes qui quittent le portefeuille de la Banque pendant la période d'observation)
4. **L'évènement par rapport auquel on souhaite étudier le temps écoulé avant la survenue:**
   - Le défaut du client sur sa carte de crédit

2. Structure des données et mesure du temps de survie (2)

5. Une variable dite de **Censure**, qui permet de déterminer si le défaut a eu lieu ou non, et ainsi identifier les comptes pour lesquels le temps de survie n'est pas connu avec exactitude parce que le compte:
    5. A fait défaut est survenu après la fin de la période d'analyse, ou
    6. Est perdu de vue sans faire défaut (par exemple en cas de fermeture de son compte avec la Banque)

6. Les variables candidates ou pressenties comme déterminants du temps de survie (potentielles variables explicatives)

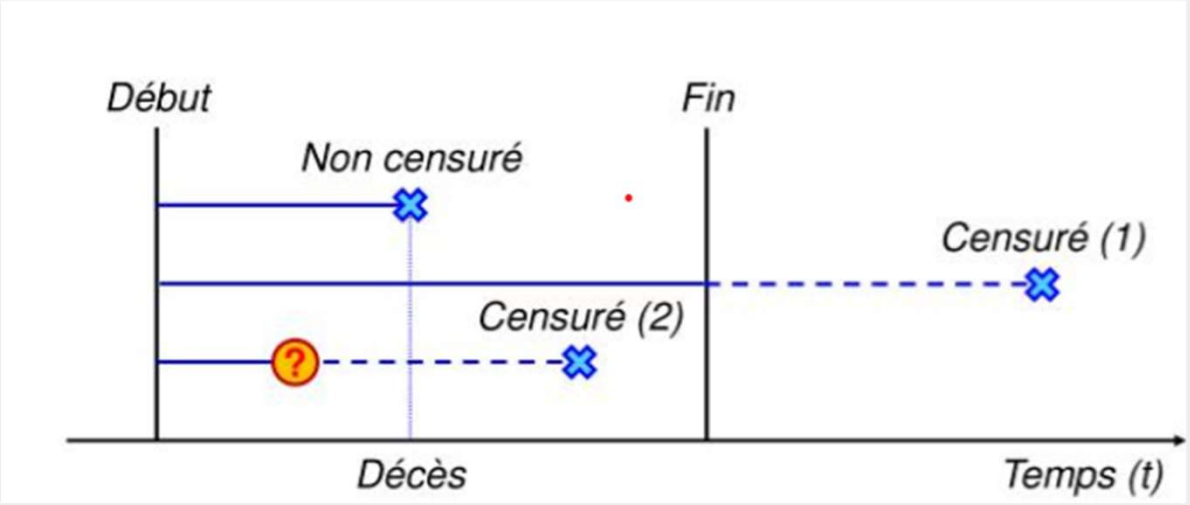
☛ **Un aperçu des 15 premières observations de notre échantillon**

ID_compte	date_observation	date_defaut	date_fermetu	censure	age	revenu	Jrs_delinquance	chomage
1	13Jan2018		13Jan2019	0	34	82 154	36	1
2	17Jan2018	16Feb2018		1	30	35 846	44	1
3	19Feb2018	20Jun2018		1	61	88 817	42	1
4	05Mar2018	04Jun2018		1	26	57 955	42	1
5	20Mar2018	19Apr2018		1	51	64 280	26	1
6	20Apr2018	20Mar2019		1	49	74 475	39	1
7	20Apr2018	19Sep2018		1	29	53 039	46	1
8	26Apr2018	24Jan2019		1	28	92 935	36	0
9	11May2018	10Jun2018		1	62	45 688	34	1
10	13May2018	10Dec2021		1	62	98 701	25	0
11	29May2018	27Nov2018		1	30	58 285	34	1
12	16Jun2018	15Apr2023		1	71	48 074	22	0
13	22Jun2018	22Jul2018		1	35	42 288	41	0
14	02Jul2018		01Aug2018	0	49	50 486	46	1
15	17Jul2018	14Feb2019		1	53	62 728	36	1

2. Structure des données et mesure du temps de survie (3)

Le temps de survie avant le défaut T est le temps le plus court écoulé entre la **date d'observation** du compte et:

1. Sa **date de fermeture** (si la carte est fermée avant la survenue du défaut) , ou
2. Sa **date de défaut** (si le défaut survient avant la date de fin de l'analyse), ou
3. La **date de fin de l'analyse** si le compte n'a ni quitté le portefeuille, ni n'est tombé en défaut avant la fin de l'analyse .



3. Formulation du modèle et sélection des variables

- On veut estimer une relation causale entre le temps de survie avant le défaut T et un ensemble de variables explicatives candidates que sont:
    1. L'âge du détenteur de la carte de crédit;
    2. Son statut d'emploi (chômeur ou non);
    3. Son revenu annuel, et
    4. Le nombre de jours de délinquance observé sur la carte de crédit.
- En modélisation du risque, on utilise très souvent des équations similaires à la régression linéaire classique.
- Dans ladite approche, le logarithme (népérien) du temps de survie Log(T) est une fonction des variables explicatives selon la formulation générale :

$$Log(T) = \underbrace{\beta_0 + \sum_{k=1}^{4} \beta_4 X_4}_{\mu} + \sigma \times \varepsilon \; (1)$$

- μ est appelé paramètre de position (« *location* » *parameter*);
- σ est appelé paramètre d'échelle (« *scale* » *parameter*);
- ε est le terme d'erreur aléatoire (indépendant des variables explicatives X_k .

3. Formulation du modèle et sélection des variables (2)

- Les β_k représentent l'influence de chaque variable explicative X_k sur le temps de survie: d'autant plus importante que β_k sera élevé.
- Si on suppose la distribution de probabilité de ε connue, celle du temps de survie T l'est également (**approche paramétrique**)
- Une transformation de l'équation (1) permet d'obtenir une forme exponentielle comme suit:

$$T = \mathrm{T}_0 \times e^{(\beta_0 + \sum_{k=1}^{4} \beta_k X_k)}, \text{ avec } T_0 = e^{\sigma \times \varepsilon}$$

- Cette expression montre que le rôle des variables explicatives est d'accélérer ou de ralentir le temps de survie: c'est pourquoi on parle de **modèle à temps accéléré** (« *Accelerated Failure Time (AFT) model*»).
- Ce type de modèle suppose donc que la fonction de survie S(t), conditionnée par les variables explicatives, se ramène à une fonction de survie de base $S_0(t)$ selon une relation: $S(t/X) = S_0\left[te^{\beta X}\right]$
  - X désigne le vecteur des variables explicatives et β celui de leurs coefficients dans le modèle, et S_0 une distribution de survie connue.
- Les lois de probabilité les plus usuelles dans la famille des modèles à temps accéléré sont: **la loi exponentielle, la loi de Weibull, la loi lognormale et la loi log-logistique.**

3. Formulation du modèle et sélection des variables (3)

- La sélection des variables les plus pertinentes à retenir parmi les variables candidates sera généralement guidée par la connaissance du secteur.
- Elle peut être manuelle lorsque la liste des variables candidates est limitée, ou automatisée lorsque ladite liste est volumineuse.
- Les logiciels statistiques intègrent parfois des algorithmes permettant la mise en œuvre d'approches de sélection automatisée de variables sur données volumineuses (approches forward, backward ou stepwise):
  - ***Forward* (ou sélection progressive):** la procédure consiste à intégrer progressivement les variables une à une, en partant du modèle avec la seule constante. La procédure s'arrête lorsque le modèle cesse de s'améliorer.
  - ***Backward* (élimination progressive):** repose sur le même principe que la procédure forward, sauf que cette fois on part du modèle complet et on élimine pas à pas les variables les moins significatives.
  - ***Stepwise*:** du même type que forward, sauf qu'il est possible que des variables introduites à une certaine étape soient retirées du sous-modèle dans une autre étape. Elle combine donc procédure forward et procédure backward.

4. Résultats de la Modélisation sous SAS

Parmi les logiciels statistiques, SAS est reconnu pour être le meilleur pour la modélisation des données de survie

- **Paramétrage des données et calcul du temps de survie**

```
%let date_fin="31DEC2023"d;

data Analyse_survie;
    set surv.Simul_survie;

    /* variable de censure, indiquant l'occurence du defaut sur la durée de vie */
    Censure=(not missing(date_defaut));

    /* Mesure du temps de survie*/
    if not missing(date_defaut) then
        Survie = intck("months",date_observation,date_defaut);
    else if not missing(date_fermeture) then
        Survie = intck("months",date_observation,min(date_fermeture,&date_fin.));
    else     Survie = intck("months",date_observation,&date_fin.);

    if survie=0 then
        delete;

    /*transformation logarithmique*/
    L_revenu=log(revenu);
run;
```

4. Résultats de la Modélisation sous SAS (2)

Mise en œuvre de la régression de survie dans SAS:

- Les procédures **PROC LIFEREG** et **PROC PHREG** du logiciel SAS permettent de modéliser les données de survie.
- La PROC LIFEREG permet de modéliser le temps de survie T.
- La PROC PHREG permet quant à elle de modéliser la probabilité instantanée en t de faire défaut sachant que l'individu a survécu jusqu'à cet instant (c'est à dire la fonction de risque h(t) décrite à la page 11).
- Les coefficients des modèles issues des deux procédures devraient donc être similaires, mais de signes contraires.
- La PROC PHREG présente l'avantage de permettre l'option de sélection automatisée des variables, utile lorsqu'on part d'une multitude de variables candidates.
- En pratique, on pourrait donc utiliser la PROC PHREG pour la sélection automatisée des variables, puis utiliser les variables retenues dans la PROC LIFEREG pour modéliser le temps de survie.
- Les variables finalement retenues dans le modèle devront combiner:
  - La significativité statistique, **et**
  - L'intuition économique (matérialisation dans les données du signe attendu).

4. Résultats de la Modélisation sous SAS (3)

Régression avec toutes les variables (PROC LIFEREG)

```
proc lifereg data= Analyse_survie;
   model survie*censure(0) = age Jrs_delinquance chomage L_revenu;
run;
```

Model Information	
Data Set	WORK.ANALYSE_SURVIE
Dependent Variable	Log(Survie)
Censoring Variable	Censure
Censoring Value(s)	0
Number of Observations	97
Noncensored Values	77
Right Censored Values	20
Left Censored Values	0
Interval Censored Values	0
Number of Parameters	6
Name of Distribution	Weibull
Log Likelihood	-118.6023266

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-1.9056	3.4152	-8.5992	4.7880	0.31	0.5769
age	1	-0.0017	0.0054	-0.0123	0.0089	0.10	0.7548
Jrs_delinquance	1	-0.0872	0.0132	-0.1131	-0.0614	43.75	<.0001
chomage	1	-1.0303	0.1844	-1.3917	-0.6690	31.24	<.0001
L_revenu	1	0.7227	0.3027	0.1294	1.3160	5.70	0.0170
Scale	1	0.7853	0.0694	0.6604	0.9339		
Weibull Shape	1	1.2734	0.1126	1.0708	1.5143		

Sélection automatisée avec la PROC PHREG

```
proc phreg data=  Analyse_survie;
   model survie*censure(0)=age Jrs_delinquance chomage L_revenu/ selection=backward;
run;
```

Step 1. Effect age is removed. The model contains the following effects:

Jrs_delinquance chomage L_revenu

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	565.319	527.426
AIC	565.319	533.426
SBC	565.319	540.457

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	37.8931	3	<.0001
Score	37.3950	3	<.0001
Wald	35.9450	3	<.0001

Note: No (additional) effects met the 0.05 level for removal from the model.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Jrs_delinquance	1	0.09597	0.01900	25.5220	<.0001	1.101
chomage	1	0.98132	0.26306	13.9161	0.0002	2.668
L_revenu	1	-0.80142	0.40117	3.9908	0.0458	0.449

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	age	1	3	0.1335	0.7148

4. Résultats de la Modélisation sous SAS (3)

```
proc lifereg data= Analyse_survie;
   model survie*censure(0) =chomage Jrs_delinquance L_revenu/distribution=exponential;
run;
```

Variable	Signe attendu	Paramètres du modèle en fonction de la distribution de probabilité (significativité entre parenthèses)			
		Log-logistique	Weibull	Exponentiel	Lognormal
Constante	n.a.	-1.2442 (0.7352)	-2.108 (0.5312)	-2.1049 (0.6199)	-1.3637 (0.7193)
Chomage	-	-0.8699 (<.0001)	-1.0193 (<.0001)	-0.9701 (<.0001)	-0.8742 (<.0001)
Jrs_delinquance	-	-0.0835 (<.0001)	-0.0871 (<.0001)	-0.0889 (<.0001)	-0.0836 (<.0001)
Revenu	+	0.6022 (0.0649)	0.7328 (0.0152)	0.7337 (0.0536)	0.6097 (0.0717)
Paramètre d'échelle σ (« scale »)		0.979	0.7848	1	0.979
Paramètre de forme δ (« shape »)		n.a.	1.2741	1	0
AIC		250.668	**247.302**	251.989	249.543

4. Résultats de la Modélisation sous SAS (4)

Choix du meilleur modèle

- Le critère Le critère AIC (*Akaike information Criterion*) permet de choisir laquelle de ces distributions convient mieux aux données utilisées
- La meilleure distribution est celle qui minimise l'AIC, ce qui revient a maximiser la vraisemblance car:

  $AIC = 2k - 2Ln(L)$, où

  - K est le nombre de paramètres estimés dans le modèle
  - L est le maximum e vraisemblance (valeur maximale de la fonction de vraisemblance du modèle)
- Selon le critère d'AIC, la **distribution de Weibull** représente mieux le temps de survie avant le défaut en utilisant nos données.

4. Résultats de la Modélisation sous SAS (5)

Modèle Final

- Selon nos données, le temps de survie T avant le défaut sur la carte de crédit serait déterminé par les retards de paiement sur la carte (Jrs délinquance), le statut à l'emploi (chômage) et le revenu du détenteur de la carte selon l'équation suivante:

$$\text{Log(T)} = \underbrace{-2.108 - 1.0193 \times \text{chômage} - 0.0871 \times \text{Jrs délinquance} + 0.7328 \times \text{Log(Revenu)}}_{Xbeta}$$

- T est distribué selon une loi de Weibull d'échelle σ=0.7848 et de forme δ=1÷ σ =1.2741
- La probabilité de défaut sur un horizon t se détermine à l'aide de la fonction de répartition de la distribution de Weibull comme suit:

$$F(T) = P(T \leq t) = e^{(-alph \times t^{\delta})} \text{ avec } Alpha = e^{(-Xbeta \times \delta)}$$

Pour un horizon de 1 an, il suffit de remplacer t par 12 mois.

Conclusion

- Originellement destinée à la recherche clinique et médicale, l'analyse de survie trouve désormais des applications dans divers autres domaines, y compris l'économie, la finance, les sciences sociales et humaines.
- En gestion des risques il devient particulièrement intéressant dans le contexte des normes IFRS9, pour estimer la probabilité de défaut sur toute la durée de vie des prêts.
- La modélisation du risque utilise généralement les approches paramétriques pour modéliser le temps de survie avant le défaut, avec les modèles à temps accéléré .
- Bien que SAS soit parmi les meilleurs logiciels recommandés pour ce faire, il pourrait s'avérer intéressant d'explorer la modélisation des données de survie a l'aide d'autres logiciels statistiques comme R, STATA, SPSS, etc..

Quelques références

1. **Allison, Paul D,** *Survival Analysis Using the SAS® System: A Practical Guide, Second Edition. Copyright © 2010,* SAS Institute Inc., Cary, NC, USA.
2. **Dachao Liu,** *PROC LIFEREG or PROC PHREG*. Northwestern University, Chicago, IL, Paper 75-2010.
3. **Joseph C. Gardiner,** *Survival Analysis: Overview of Parametric, Nonparametric and Semiparametric approaches and New Developments*. SAS Global Forum 2010 (Paper 252-2010).
4. **Lida Gharibvand,** *A Step-by-Step Guide to Survival Analysis***.** University of California, Riverside.
5. **Tony Bellotti**, *Discrete Survival Models for Retail Credit Scoring*. University of Bristol, 14 November 2014.
6. **Viani Djeundje and Jonathan Crook**, *Dynamic survival models for credit risks*. Credit Research Centre, University of Edinburgh

Merci

Josiane Guedem
ejfotso@yahoo.fr