Analyse de survie

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

Analyse de survie

Étude du temps avant qu'un événement survienne.

- temps qu'un client demeure abonné à un service offert par notre compagnie.
- ancienneté d'un travailleur au service d'une compagnie.
- durée de vie d'une franchise ou avant la faillite d'une entreprise.
- temps avant le prochain achat d'un(e) client(e).
- temps durant lequel une personne est au chômage.

Mécanismes de survie

La survie est caractérisée par la présence d'information **partielle**.

Les données sont sujettes à **troncature** et à **censure**.

Définition de la censure

Le temps réel de l'événement n'est pas observé (information partielle).

- Censure à droite: l'événement n'est pas encore survenu au temps t: on sait que $T > t$.
- Censure à gauche: l'événement survient avant le temps t, donc la vraie valeur est inférieure à la valeur observée ($T < t$)
- Censure par intervalle: l'événement est survenu dans la plage $[t_1, t_2]$ (données arrondies)

Définitions de la troncature

La plage des valeurs possibles est tronquée.

- troncature à gauche: le temps minimum est supérieur à t_0
- troncature à droite: le temps maximum est inférieur à t_1
- troncature par intervalle: le temps de l'événement doit survenir entre t_0 et t_1.

Exemple: étude sur le chômage

- Certaines personnes seront déjà au chômage au début de l'étude (troncature à gauche): le temps réel sera supérieur à la durée écoulée
- Certaines personnes ont trouvé un emploi entre deux prises de contact, mais la date exacte est inconnue (censure par intervalle).
- D'autres personnes seront toujours au chômage à la fin de l'étude (censure à droite).

Diagramme de Lexis

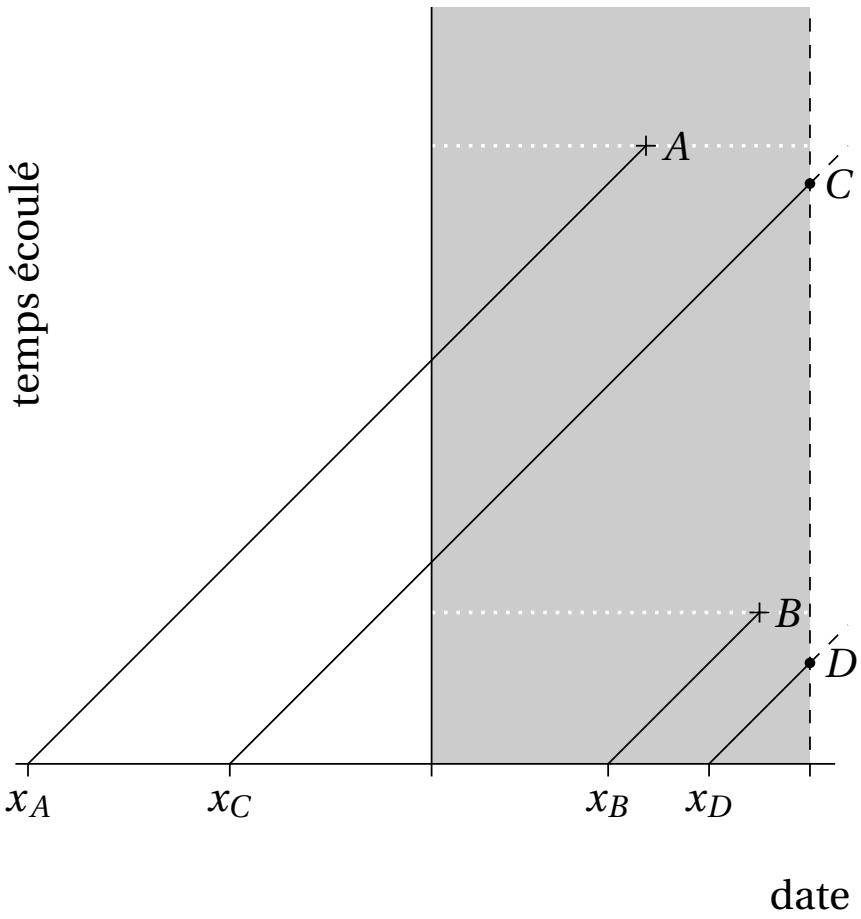
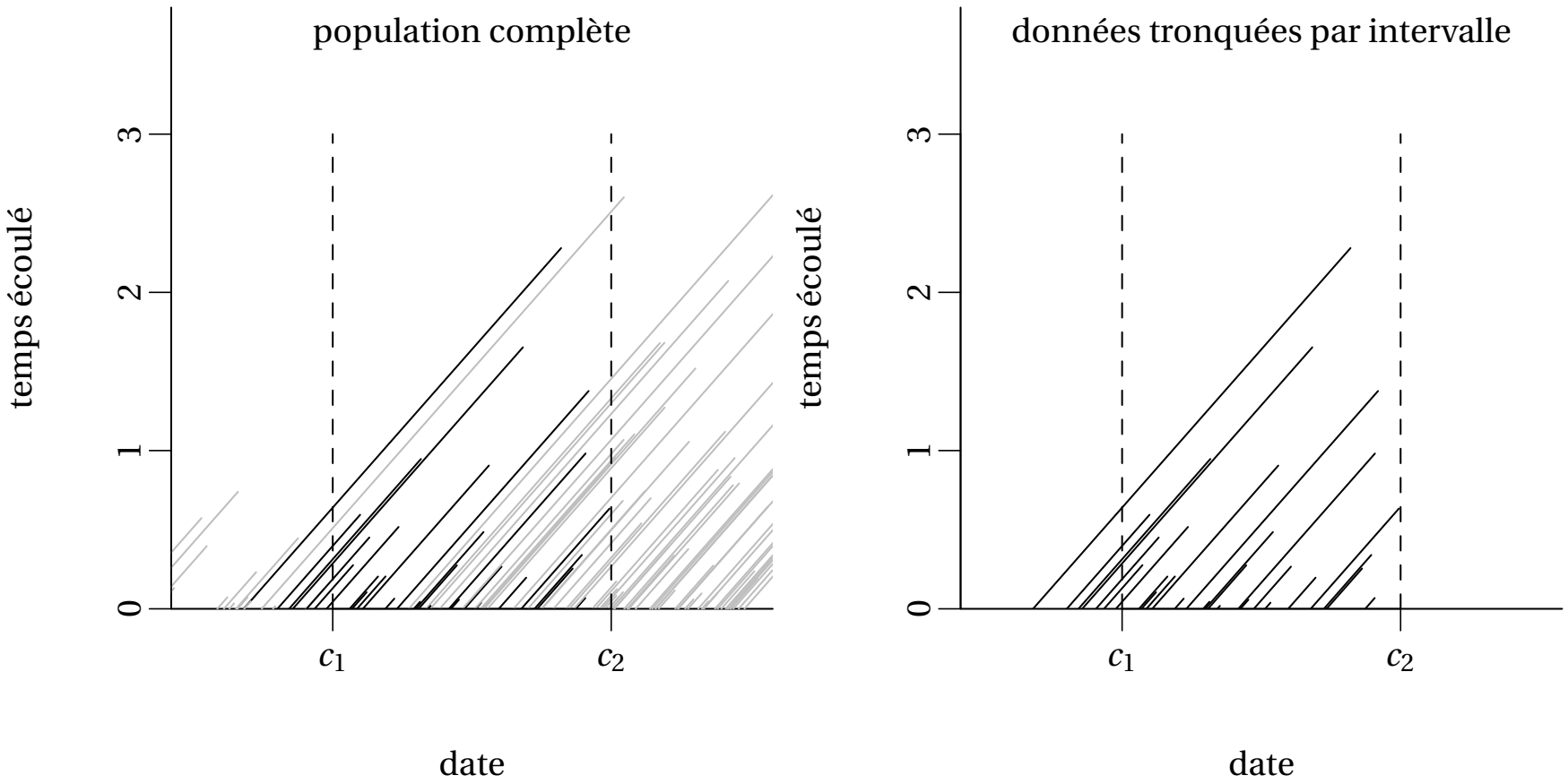


Figure 1: Diagramme de Lexis pour données tronquées à gauche (A et C) et censurées à droite (C et D).

- On trace une droite de pente 1 représentant la durée en fonction du temps (date au calendrier).
- La fenêtre définit la période de collecte de donnée.
- On peut lire le temps initial et le temps final sur l'axe des y.
- La censure est indiquée par des cercles, les événements par des croix.

Troncature par intervalle



On s'intéresse à la durée de la relation d'emploi: seules les personnes à l'emploi qui ont pris leur retraite entre 2009 et 2021 sont considérées pour l'étude.

Exemple de données de survie

Tableau extrait de Hanamaya et Sibuya (2016)

		Year of Birth												
		1875	1876	1877	1878	1879	1880	1881	1882	1883	1884	1885	1886	1887
Male	100	55	69	75	96	120	93	141	146	170	208	185	227	288
	101	32	36	42	57	67	55	83	85	102	135	115	146	185
	102	18	18	26	39	44	33	53	54	63	67	71	90	105
	103	10	7	15	23	28	21	37	30	37	39	38	58	64
	104	4	1	8	13	18	12	20	14	23	19	18	33	33
	105	2	0	5	10	9	7	9	9	15	13	6	17	15
	106	2	0	2	2	7	4	3	6	9	4	5	9	10
	107	2	0	1	0	2	3	2	5	3	1	3	7	7
	108	1	0	1	0	2	1	2	2	1	1	1	5	2
	109	1	0	0	0	1	1	1	0	0	1	0	3	2
	110	1	0	0	0	1	0	0	0	0	1	0	1	1
	111	1	0	0	0	0	0	0	0	0	1	0	1	0
	112	0	0	0	0	0	0	0	0	0	1	0	1	0
	113	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2: Âge au décès (cohortes éteintes) de centenaires Japonais.

Données censurées par intervalles et tronquées à droite (l'âge maximum des personnes en date de collecte des données en 2012).

Censure informative ou pas?

La censure peut être aléatoire ou non-informative:

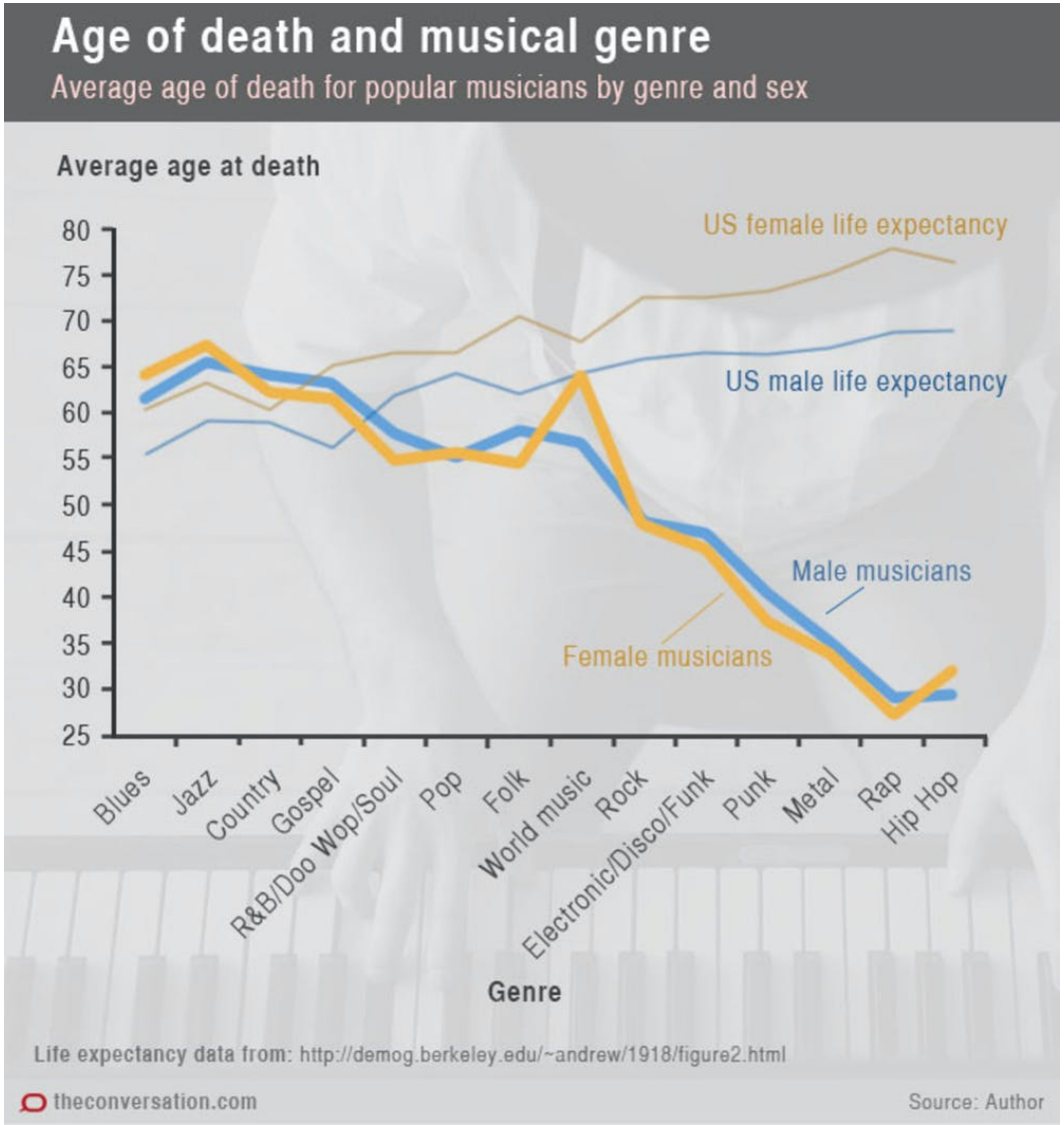
- une étude finit en juin (censure administrative)
- une personne déménage dans une autre province et n'est plus suivie

Si la censure est informative, les outils présentés ne sont pas adéquats!

- un patient d'une étude clinique est déchargé d'un protocole médical car il est trop mal en point.

Exemple

Où est l'erreur de logique dans l'analyse suivante?



Contenu du cours

La survie est un sujet très compliqué...

On s'intéressera uniquement à la censure à droite (non informative).

Le cours ne couvrira que des méthodes nonparamétriques ou semiparamétriques.

Données

La structure de données de base que l'on doit avoir pour travailler est la suivante:

1. une variable temps, T.
2. une variable binaire C (censure).
3. des variables explicatives $X_1, \dots, X_p$

On considère l'exemple d'une entreprise de télécommunications qui veut connaître les facteurs influençant la durée d'abonnement à son service de téléphone cellulaire.

Temps d'abonnement à un forfait cellulaire

Les données `survie1` contiennent les variables suivantes.

- `temps`: temps (en semaines) que le client est resté abonné au service de téléphone cellulaire. Il s'agit du vrai temps si le client n'est plus abonné et d'une borne inférieure si le client est toujours abonné.
- `censure`: variable binaire qui indique si la variable `t` est censurée (0 si le client est toujours abonné) ou non (1, la variable `t` est la durée finale de l'abonnement).
- `age`: âge du client au début de l'abonnement.
- `sexe`: sexe du client, soit femme (1), soit homme (0).
- `service`: nombre de services en plus du cellulaire auquel le client est abonné parmi internet, téléphone fixe, télévision (câble ou antenne parabolique).
- `region`: région où habite le client en ce moment (valeurs entre 1 et 5).

Survie vs régression

réponse Y	résumé descriptif	comparaison de deux groupes	modèle général
continue	moyenne	test-t pour deux échantillons	régression linéaire
binaire	proportion	test d'indépendance du khi-deux	régression logistique
temps de survie censuré à droite	fonction de survie, temps de survie médian	test log-rang, test de Wilcoxon généralisé (Gehan)	modèle de Cox

Fonction de survie

Soit $F(t) = \Pr(T \leq t)$ la fonction de répartition du temps de survie t.
La fonction de survie,

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

donne la probabilité que le temps de survie soit supérieur à t.

Fonction de risque

La **fonction de risque** (en anglais, *hazard*) est

$$h(t) = \lim_{\mathrm{d}t \to 0} \frac{\Pr(T \in [t, t + \mathrm{d}t) \mid T > t)}{\mathrm{d}t} = \frac{f(t)}{S(t)}$$

où $f(t)$ est la fonction de densité (pour T continu) ou de masse pour T discret.

Plus le risque est élevé au temps t, plus l'événement est susceptible d'arriver.

Fonction de risque (bis)

Dans le cas discret où le temps peut seulement prendre les valeurs $0, 1, 2, ...$, le risque est

$$h(t) = \frac{\Pr(T = t)}{\Pr(T \geq t)},$$

la probabilité conditionnelle que l'événement survienne au temps t, étant donné qu'il n'était pas survenu juste avant.

Risque vs survie

Les fonctions de survie et de risque sont intimement reliées, avec

$$h(t) = -\frac{\mathsf{d}\ln\{S(t)\}}{\mathsf{d}t}, \qquad S(t) = \exp\left\{-\int_0^t h(u)\mathsf{d}u\right\}.$$

Cette équation sert à illustrer qu'un modèle pour la fonction de survie spécifie une fonction de risque (et vice-versa).

Un profil avec un risque plus élevé (pour tout temps t) aura une survie plus courte.

Estimateur de Kaplan–Meier

L'estimateur de Kaplan–Meier est un estimateur de la fonction de survie en présense de censure à droite.

Cette méthode est nonparamétrique en ce sens qu'on ne suppose aucun modèle et qu'on suppose uniquement que la censure est non-informative.

Kaplan–Meier avec **R**

```r
library(survival)
data(survie1, package = "hecmulti")
# Estimateur de Kaplan-Meier
# La réponse "temps" est le temps de survie
# et l'indicateur de censure "censure" est
# "0" pour censuré à droite, "1" pour événement
kapm <- survfit(
  Surv(temps, censure) ~ 1,
  #~1 => aucune variable explicative
  type = "kaplan-meier",
  conf.type = "log", #type d'intervalle de conf.
  data = survie1)
```

Méthodes pour `survfit`

```r
# Tableau résumé de la survie
summary(kapm)
# Graphique de la fonction de survie
plot(kapm) # graphique de base
# Quantiles (par défaut, quartiles)
quantile(kapm)
```

Parmi les 500 observations, il y a

- 334 clients qui ont terminé leur abonnement et
- 166 qui sont censurées à droite.

La fonction résumé (`summary`) renvoit l'estimation de la fonction de survie pour chaque temps d'échec (événement).

- l'estimateur est *indéfini* à ces valeurs.

Tableau résumé (sortie tronqué)

temps	nb à risque	nb échecs	nb cumul.	survie	erreur-type
2	500	1	1	0.9980	0.002
11	499	1	2	0.9960	0.003
14	498	1	3	0.9940	0.003
18	497	1	4	0.9920	0.004
27	496	1	5	0.9900	0.004
29	495	1	6	0.9880	0.005
30	494	1	7	0.9860	0.005
34	493	4	11	0.9780	0.007
189	13	1	331	0.2037	0.028
202	6	1	332	0.1697	0.039
216	2	2	334	0.0000	

La probabilité estimée que le temps d'abonnement dépasse 30 semaines est 0.986.

Graphique de la fonction de survie

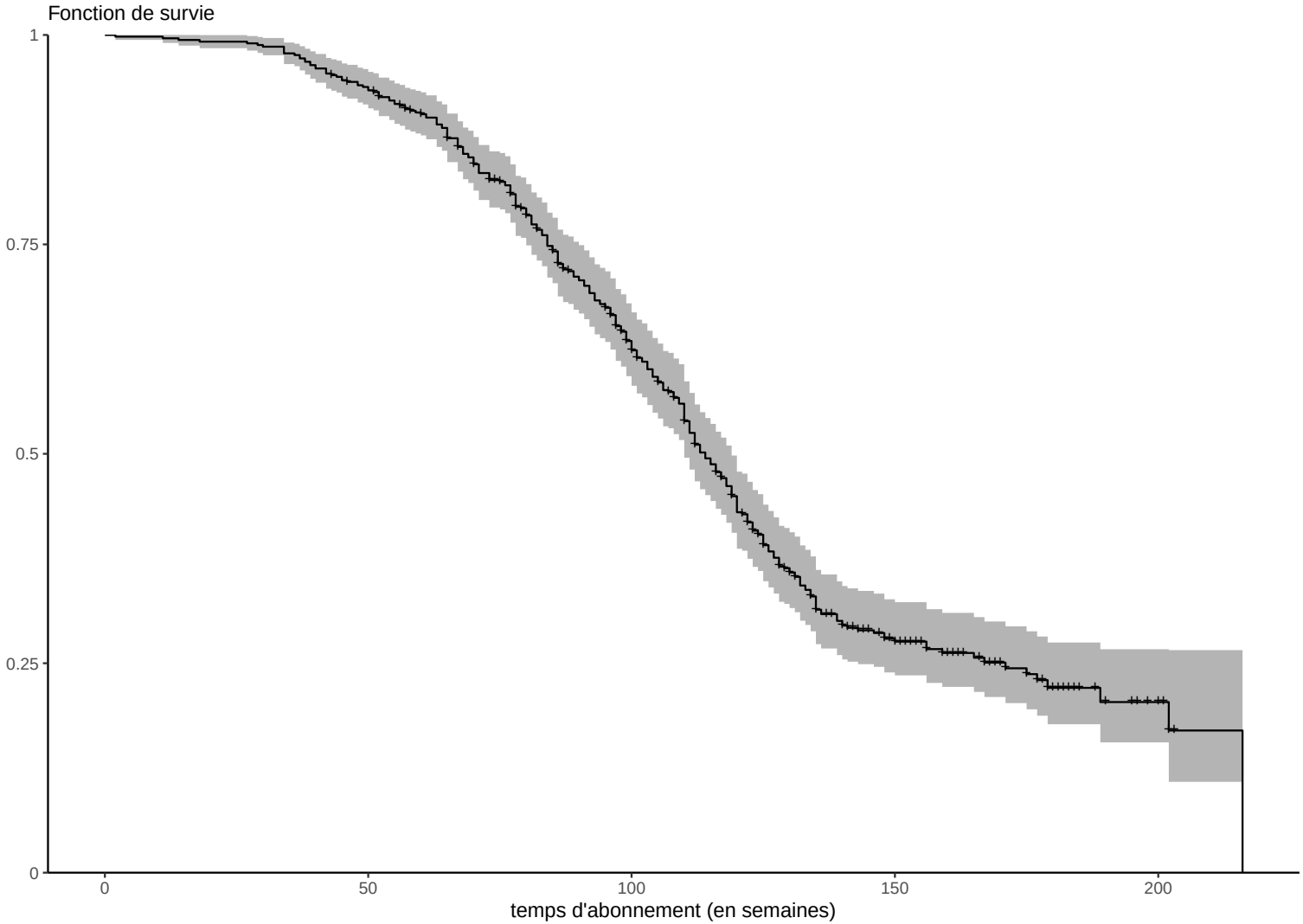


Figure 3: Estimation de Kaplan–Meier de la fonction de survie pour les données d'abonnement avec intervalles de confiance ponctuels à 95%.

Comprendre l'estimateur de Kaplan–Meier

- On parle d'**échec** au temps t_i si $T = t_i$ et $C = 0$ (événement observé au temps t_i).
- Le nombre de **personnes à risque** au temps t_i est le total des observations dont le temps mesuré excède t_i (censure et événements postérieurs à t_i)

Construction:

- Ordonner les temps (uniques) où il y a des échecs (temps où `censure = 1`), disons $t_{(1)} \leq \cdots \leq t_{(m)}$
- À chaque temps $t_{(i)}$ $(i = 1, \ldots m)$, on calcule le nombre de personnes à risque, r_i, et le nombre d'échecs, d_i.
- Le risque empirique est $\widehat{h}_i = r_i/d_i$, la proportion d'échecs parmi les personnes à risque.

Dérivation de Kaplan–Meier

L'estimateur de Kaplan–Meier définit une **fonction escalier**

- Entre $t = 0$ et $t = t_{(1)}$, la survie est de 1.
- Entre $t = t_{(1)}$ et $t = t_{(2)}$, la survie est $1 - \widehat{h}_1$.
- Entre $t = t_{(2)}$ et $t = t_{(3)}$, la survie est $(1 - \widehat{h}_1) \times (1 - \widehat{h}_2)$.

Pour un temps t donné, on multiplie tous les termes $(1 - \hat{h}_i)$ des temps d'échecs passés,

$$\hat{S}(t) = \prod_{i:t_{(i)}<t} \left(1 - \hat{h}_i\right).$$

- la survie ne change qu'aux valeurs de $t_{(i)}$ $(i = 1, \dots, m)$
- les contre-marches interviennent uniquement aux temps observés d'**échecs**.
- si la plus grande observation est censurée, la courbe de survie n'atteindra jamais zéro.

Quantiles de la fonction de survie

On obtient le quantile à partir du tableau: c'est le premier temps d'échec où la survie est inférieure au quantile (la ligne horizontale au quantile traverse une contre-marche).

niveau (%)	quantile	IC 95% (borne inf.)	IC 95% (borne sup.)
25	84	80	90
50	114	110	119
75	171	143	

- L'estimé du temps de survie médian est de 114 semaines: on estime que la moitié des clients vont avoir une durée d'abonnement supérieure (ou inférieure) à 114 semaines.
- Un intervalle de confiance (IC) de niveau 95%, pour le temps médian est [110, 119] semaines (IC de Wald, transformé).

Estimateur déficient et moyenne restreinte

- Si la plus grande observation n'est pas censurée à droite, il est possible d'estimer la moyenne.
  - aire sous la courbe
  - autrement, on obtient une borne inférieure (sous-estimation de la moyenne), appelée *moyenne restreinte*.

```
1 print(kapm, print.rmean = TRUE)
```

- La moyenne estimée via Kaplan–Meier est 125 semaines (`rmean`).
- Les statistiques descriptives usuelles sont à proscrire! Elles ne tiennent pas compte de la censure.
  - Par exemple, la moyenne empirique ici est de 107.788 semaines.

Comparaison de plusieurs courbes de survie

Si on sépare l'échantillon selon une variable catégorielle en K groupes, on peut estimer séparément les fonctions de survie des groupes, disons $S_1(t), \dots, S_K(t)$.

On peut tester l'égalité des fonctions de survie, c'est-à-dire, les hypothèses

- $\mathscr{H}_0 : S_1(t) = \cdots = S_K(t)$ pour tout t et
- $\mathscr{H}_a$: les courbe de survies diffèrent pour au moins une valeur de t.

Les deux tests utilisés habituellement sont

- le test du log-rang et
- le test de Wilcoxon généralisé (ou test de Gehan).

Comparaison de la survie selon le sexe

On stratifie l'échantillon selon le sexe.

```
1 survdiff(formula = Surv(temps, censure) ~ sexe,
2           data = survie1)
```

```
Call:
survdiff(formula = Surv(temps, censure) ~ sexe, data = survi

        N Observed Expected (O-E)^2/E (O-E)^2/V
sexe=0 309      217      181      7.33      16.4
sexe=1 191      117      153      8.63      16.4

 Chisq= 16.4  on 1 degrees of freedom, p= 5e-05
```

- La fonction `survdiff` avec la formule retourne le résultat du test asymptotique du log-rang pour l'hypothèse d'égalité des fonctions de survie.
- La valeur-p du test est inférieure à 10^{-4}: on rejette $\mathscr{H}_0$ pour conclure qu'il y a une différence significative entre les deux courbes de survie.

Courbes de survie selon le sexe

On voit que la courbe des femmes est systématiquement au-dessus de celle des hommes. Les femmes ont donc tendance à rester abonnées plus longtemps que les hommes.

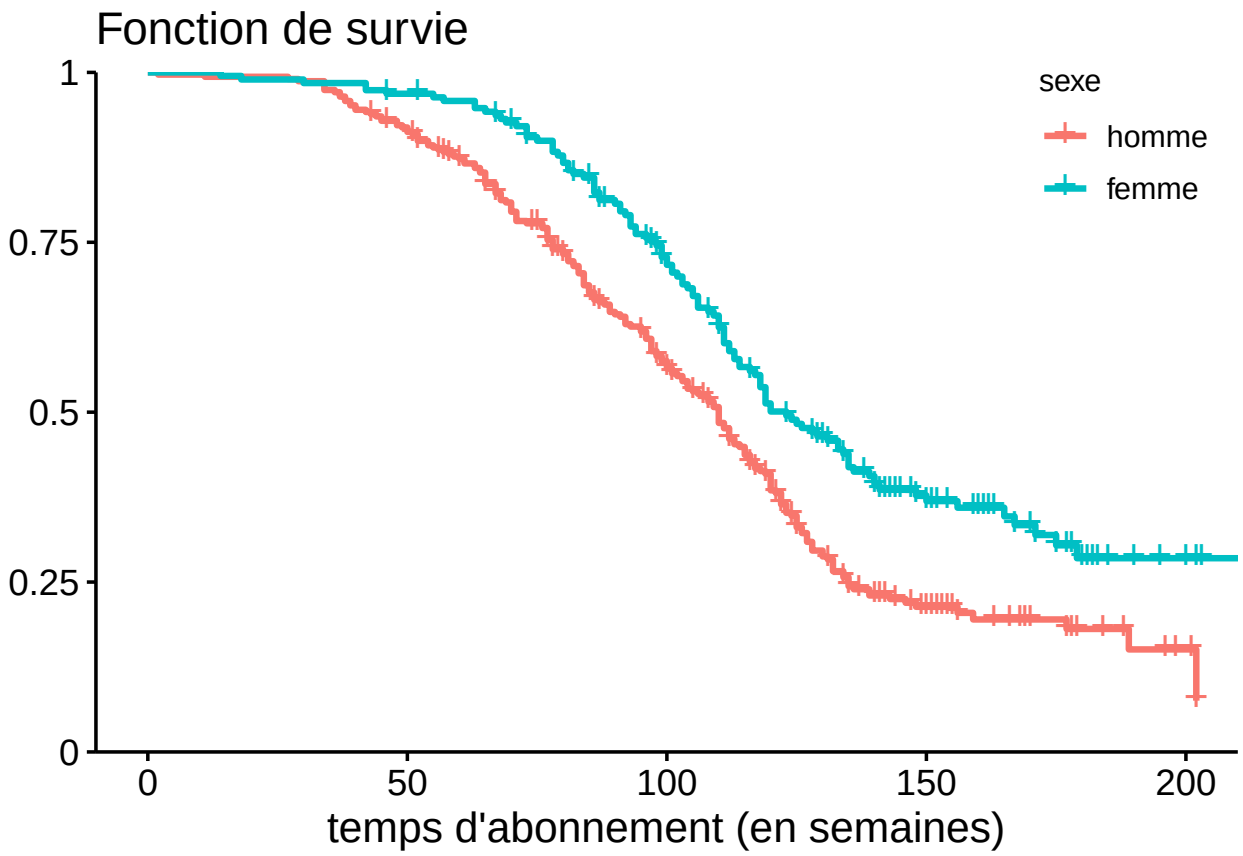


Figure 4: Courbes de survie pour les durées d'abonnement selon le sexe.

Modélisation par groupe

L'estimateur de Kaplan–Meier ne permet pas l'inclusion de variables explicatives à proprement parler.

- On peut diviser l'échantillon en sous-groupes (stratification).
- On utilise ensuite l'estimateur de Kaplan–Meier pour chacune des modalités.
- Cela réduit la taille de l'échantillon disponible et rend l'estimation plus incertaine.

Récapitulatif

- L'analyse de survie est l'étude de temps d'attente (variable positive) avant que survienne un événement.
- L'étude des temps de défaillance nécessite l'utilisation d'outils statistiques spécifiques en raison des mécanismes de censure et de troncature.

Récapitulatif

- La fonction de survie $S(t)$ encode la probabilité que le temps de défaillance excède le temps t.
- La fonction de risque encode la probabilité de mourir juste après le temps t sachant qu'on a survécu jusque là.
- La connaissance de la fonction de survie permet d'obtenir la fonction de risque et vice-versa.

Récapitulatif

- Le mécanisme d'information partielle le plus courant en analyse de survie est la **censure à droite** (on ne connaît qu'une borne inférieure pour le temps de défaillance, l'événement n'étant toujours pas survenu au temps donné).
- Si on traitait les temps de censure comme des temps de défaillance observée, on *sous-estimerait* la durée de survie.

Récapitulatif

- L'estimateur de Kaplan–Meier est l'estimateur du maximum de vraisemblance nonparamétrique si on a de la censure à droite aléatoire ou non-informative. Il ne fait aucun postulat sur la distribution de la survie.
- Pour que l'estimation soit de qualité, il faut un nombre *conséquent* d'observations.
- La quantité d'observations censurées impacte la précision de l'estimation.
- L'estimateur est déficient si le plus grand temps observé est censuré à droite (l'estimation de la fonction de survie ne décroît pas à zéro)

Récapitulatif

- Le test du log-rang permet de valider si plusieurs fonctions de survie sont égales (en tout temps).
- On peut estimer la fonction de survie indépendamment pour chaque modalité d'une variable explicative catégorielle en stratifiant: cela réduit la taille de l'échantillon dans chaque strate.

Modèle à risques proportionnels de Cox

Si on veut inclure l'effet de variable explicatives, on se tourne vers le **modèle à risques proportionnels de Cox**.

Ce dernier spécifie que le risque au temps t est

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

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

- Le postulat de risques proportionnels implique que le terme $\exp(\mathbf{X}\beta)$ — et donc les coefficients β — ne varient pas selon le temps t.
- La fonction $h_0(t)$ peut être interprétée comme la fonction de risque lorsque toutes les variables explicatives $\mathbf{X}$ valent zéro.

Composante paramétrique

Le terme $\exp(\mathbf{X}\beta)$ modélise l'effet d'un changement des valeurs des variables explicatives sur la fonction de risque de base.

Le modèle de Cox est un modèle multiplicatif:

- Si X_i augmente d'une unité le rapport de risque est de $\exp(\beta_i)$.
- Pour chaque augmentation d'une unité de X_i, le risque que l'événement survienne est multiplié par $\exp(\beta)$, *ceteris paribus*.

Modèle de Cox dans **R**

Le modèle de Cox, `coxph`, inclut un objet de classe `Surv()` avec le temps de défaillance et l'indicateur de censure comme variable réponse.

Il y a trois options pour la gestion des doublons: `ties = "exact"` est meilleur, mais plus coûteux (parfois prohibitivement).

```
cox <- coxph(
  Surv(temps, censure) ~
    age + sexe + region + service,
  data = survie1,
  ties = "exact") # gestion des doublons
# Tableau résumé avec coefficients,
# intervalles de confiance de Wald
# et tests pour significativité globale
summary(cox)
# Test du rapport de vraisemblance
car::Anova(cox, type = 3)
```

Coefficients du modèle

Table 2: Rapport de risque et intervalles de confiance à niveau 95% de Wald pour le modèle de Cox.

terme	exp(coef)	borne inf.	borne sup.
age	0.950	0.937	0.962
sexe	0.511	0.404	0.646
region2	0.674	0.465	0.978
region3	1.030	0.729	1.457
region4	0.799	0.566	1.128
region5	0.966	0.683	1.366
service1	0.353	0.275	0.453
service2	0.174	0.120	0.251
service3	0.115	0.070	0.190

Ceteris paribus, le risque qu'une femme (`sexe = 1`) se désabonne est 0.511 fois plus petit que celui d'un homme (`sexe = 0`).

Tests du rapport de vraisemblance

Les statistiques de rapport de vraisemblance sont comparées à une loi khi-deux (χ^2) avec `ddl` degrés de liberté.

Table 3: Tests du rapport de vraisemblance pour les effets de type 3.

terme	statistique	ddl	valeur-p
age	68.07	1	<1e-04
sexe	32.82	1	<1e-04
region	7.67	4	0.1
service	159.31	3	<1e-04

Par exemple, l'effet marginal (une fois que les autres variables sont incluses) de la variable sexe est significatif (valeur-p inférieure à 10^{-4}).

La variable `region` n'est pas globalement significative.

Prédiction de la survie

On peut passer une base de données, ici `survie2` à un modèle de type `coxph` pour prédire la survie.

```
# Ajuster un modèle avec deux variables
data(survie2, package = "hecmulti")
cox <- coxph(
  Surv(temps, censure) ~
    age + sexe,
  data = survie1)
pred <- survfit(
  cox,        # Modèle de Cox
  newdata = survie2, # nouvelle base de données
  type = "kaplan-meier") # survie
plot(pred) # graphe de base
```

Profil des nouveaux clients

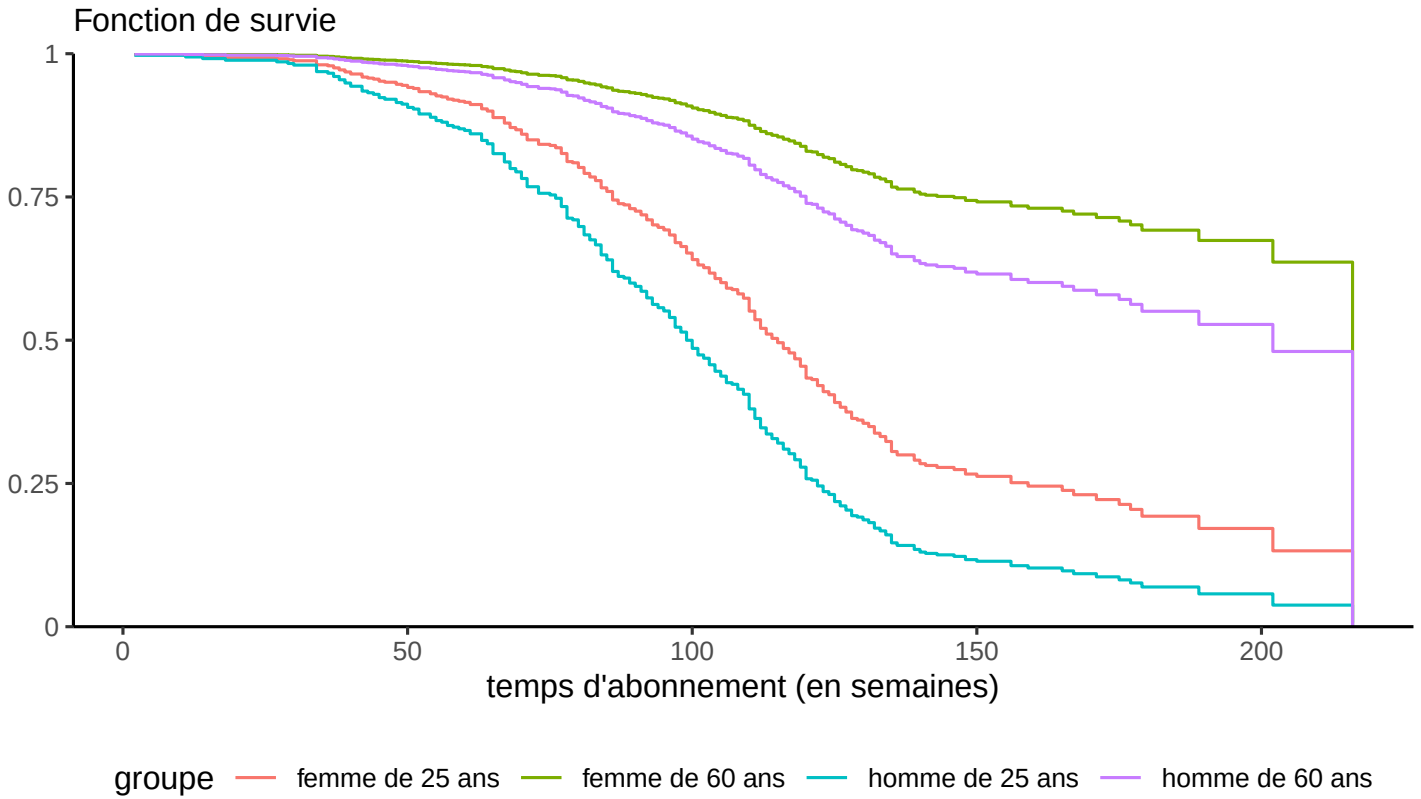
Table 4: Information sur les coefficients et les profils clients.

terme	exp(coef)	borne inf.	borne sup.
age	0.958	0.945	0.970
sexe	0.614	0.489	0.773

Table 5: Profil de quatre clients pour la prédiction.

sexe	age
0	25
1	25
0	60
1	60

Graphique des fonctions de survie



Postulat de risques proportionnels

Il est possible de créer des résidus du modèles et de faire des graphiques diagnostics pour potentiellement infirmer le postulat de risques proportionnels.

Si l'hypothèse tient la route, alors il ne devrait pas y avoir de tendance temporelle dans les résidus.

La commande `cox.zph` permet de tester le postulat de risques proportionnels à l'aide d'un test du score pour voir si la pente $\beta(t)$ associée à une covariable est nulle en fonction du temps t.

```
test_score_rprop <- cox.zph(cox)
test_score_rprop
plot(test_score_rprop)
```

Graphique des résidus

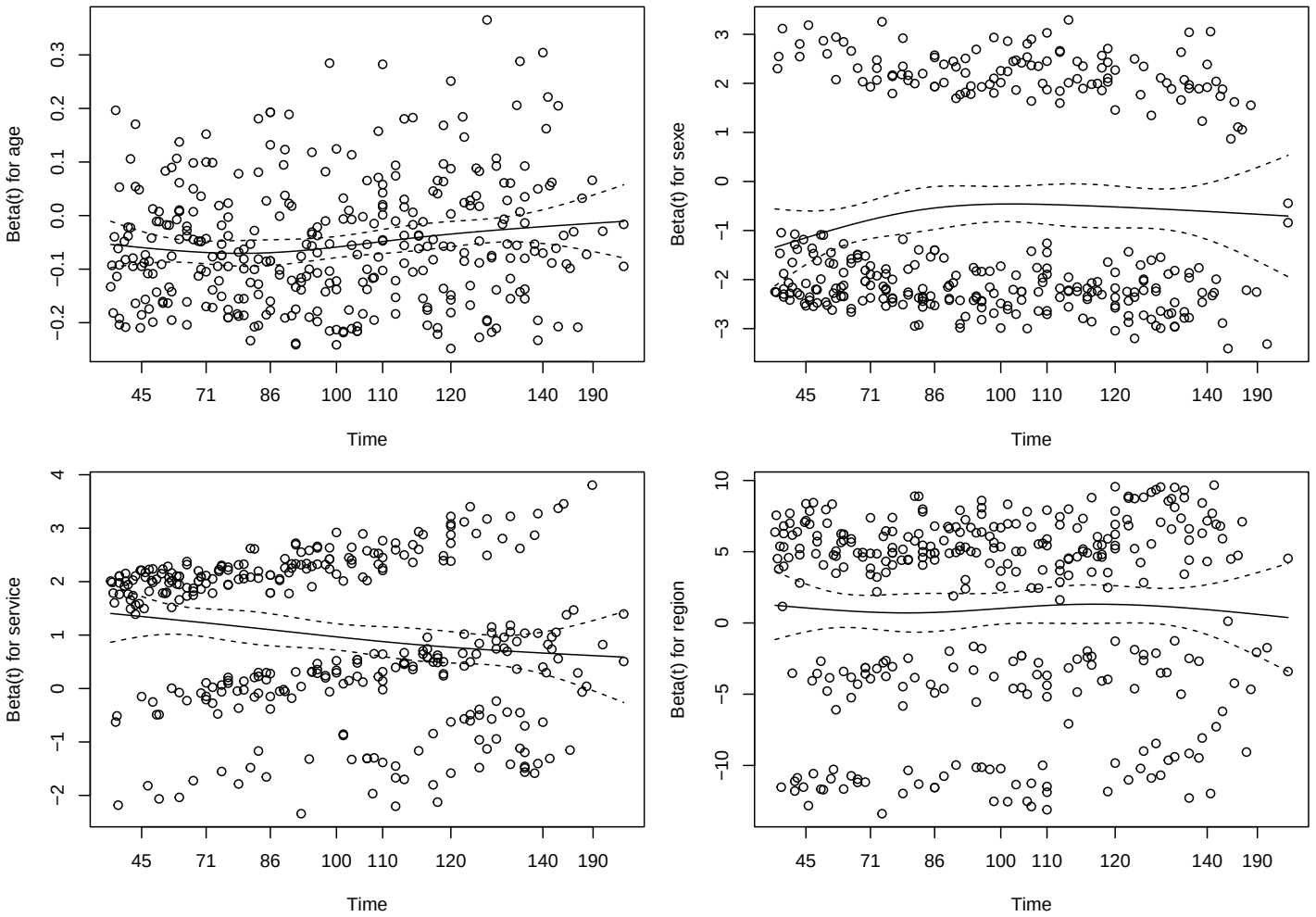


Figure 5: Estimations des coefficients en fonction du temps basés sur les moindres carrés pondérés (diagnostic graphique de Grampsch et Therneau).

Test du score pour risques proportionnels

Test du score de Grampsch et Therneau (1994) pour les coefficients constants dans le temps avec

- $\mathscr{H}_0$: $\beta(t) = c$ est constant (risques proportionnels)
- $\mathscr{H}_a$: $\beta(t) \neq c$ (risques non proportionnels, coefficient variable).

Table 6: Tests du score avec valeur-p basée sur la loi nulle khi-deux.

effet	score	ddl	valeur-p
age	4.17	1	0.041
sexe	1.09	1	0.296
service	10.62	3	0.014
region	3.81	4	0.432
global	20.83	9	0.013

Postulat pas respecté pour `service`: augmentation au fil du temps pour tous les groupes. Il faudrait ajouter une interaction ou stratifier par `service`.

Postulat de risques proportionnels

Si le postulat n'est pas validé, on peut interpréter l'effet comme un rapport de risque moyen pondéré sur la période de suivi, mais ce dernier change selon le moment (Stensrud et Hernan, 2020).

Cela implique surtout que les erreurs-types associées aux estimations sont trompeuses.

Récapitulatif

Le modèle de Kaplan–Meier ne permet pas d'estimer l'impact de variables explicatives sur la survie.

On peut utiliser pour ce faire le modèle de Cox.

Ce dernier suppose qu'on peut diviser le risque en deux parties

- risque de base $h_0(t)$ commun à tou(te)s (composante nonparamétrique)
- effet multiplicatif $\exp(\mathbf{X}\beta)$ (composante paramétrique)

Récapitulatif

Puisque $h_0(t)$ est commune à toutes les observations, moins d'incertitude sur l'estimation de la survie.

- L'impact sur la survie de changement dans les variables explicatives n'est pas multiplicatif!
- voir les prédictions pour les quatre profils clients.

Le modèle de Cox suppose que le rapport de cote ne dépend pas du temps (postulat de risques proportionnels).

- On peut vérifier ce postulat
- et généraliser le modèle au besoin.