Econométrie des modèles de durée

Guillaume Horny*

*Banque de France et UCLouvain

Master 2 MOSEF

Introduction

Pourquoi modéliser des durées?

- pour étudier la durée de séjour dans un état...
- ... mais aussi l'état suivant, vers lequel la transition a eut lieu

Exemple d'applications:
mortalité, fiabilité d'un équipement, chômage, défaillance d'entreprises, cotations d'entreprise, prévisions météorologiques...

Exemple: données

	time	status	x
1	9	1	Maintained
2	13	1	Maintained
3	13	0	Maintained
4	18	1	Maintained
5	23	1	Maintained
6	28	0	Maintained
7	31	1	Maintained
8	34	1	Maintained
9	45	0	Maintained
10	48	1	Maintained
11	161	0	Maintained
12	5	1	Nonmaintained
13	5	1	Nonmaintained
14	8	1	Nonmaintained
15	8	1	Nonmaintained
16	12	1	Nonmaintained
17	16	0	Nonmaintained
18	23	1	Nonmaintained
19	27	1	Nonmaintained
20	30	1	Nonmaintained
21	33	1	Nonmaintained
22	43	1	Nonmaintained
23	45	1	Nonmaintained

Source : R. Miller (1997), *Survival Analysis*, John Wiley & Sons.

Spécificité des durées

- VA strictement positives
- le temps est irréversible
- dépendance par rapport au temps
- données censurées
- hétérogénéité non-observée
- distributions des durées souvent non-centrées, très asymétriques
- plusieurs types de destinations possibles

Exemple: durées de chômage
$\Rightarrow$ techniques économétriques spécifiques

Amorce de bibliographie

Gouriéroux, C. (1989) : “Economérie des variables qualitatives”, Economica.
Cameron, C. et Trivedi, P. (2005): “Microeconometrics”
La plupart des manuels comprennent un chapitre sur les durées...

Pour aller plus loin:
Lancaster, T. (1991): “The Econometric Analysis of Transition Data”, Cambridge University Press.
Therneau, T. et Grambsch, P. (2000) : “Modeling Survival Data”, Springer.

Chapitre 1

Chapitre 1

Outils de base

Chapitre 1

Intuition

Du temps passé dans un état à la probabilité d'en sortir...

Notations

Variable d'intérêt: Soit T une variable aléatoire continue, positive, représentant la durée passée dans un état

Caractérisation de la loi de T

La distribution de T est entièrement caractérisée par l'une des fonctions:

- fonction de répartition $F(t)$
- fonction de densité de probabilité $f(t)$
- fonction de survie
- fonction de hasard
- fonction de hasard intégré

Fonction de répartition

$$F(t) = \Pr(T \leq t), t > 0 \tag{1}$$

Propriétés:

- $F(t) \in [0, 1], t > 0$
- F monotone non-décroissante : $t_1 < t_2 \rightarrow F(t_1) \leq F(t_2)$
- $F(0_+) = 0$ et $\lim_{t \rightarrow \infty} = 1$

Densité de probabilité

$$f(t) = \frac{dF}{dt} = \lim_{dt \to 0} \frac{\Pr(T \in [t, t + dt[)}{dt} \tag{2}$$

ou encore:

$$F(t) = \int_0^t f(u)du \tag{3}$$

Propriétés:

- f est positive car F est non-décroissante
- $\int_0^{+\infty} f(u)du = 1$
- $\int_0^a f(u)du = \Pr(T < a) = F(a)$

Fonction de survie

$$S(t) = \Pr(T > t) = 1 - F(t) \quad (4)$$

Ou encore:

$$S(t) = \int_t^{\infty} f(u)du = 1 - \int_0^t f(u)du \quad (5)$$

Propriétés:

- $S(t) \in [0, 1], t > 0$
- S est monotone non-croissante : $t_1 < t_2 \rightarrow S(t_1) \geq S(t_2)$
- $S(0_+) = 1$ et $\lim_{t \rightarrow \infty} = 0$ (l'événement d'intérêt est possible, même s'il ne se produit qu'à la fin des temps)

Fonction de hasard

$$\lambda(t) = \lim_{dt \to 0} \frac{1}{dt} P(T \in [t, t + dt[| T \geq t). \tag{6}$$

Interprétation:
Probabilité instantanée (ni une densité, ni une probabilité) de quitter à la date t l'état actuel, sachant que l'individu est dans cet état en t.
Démographie: le taux de mortalité correspond à $\lambda(t)dt$, car pour dt petit:

$$\lambda(t)dt \approx P(T \in [t, t + dt[| T \geq t). \tag{7}$$

Fonction de hasard (suite)

On a:

$$\begin{aligned}\lambda(t) &= \lim_{dt\to 0} \frac{1}{dt} \frac{P((t \leq T < t+dt) \cap (T \geq t))}{\Pr(T \geq t)} \\ &= \lim_{dt\to 0} \frac{1}{dt} \frac{P(t \leq T < t+dt)}{\Pr(T \geq t)} \\ &= \frac{f(t)}{S(t)}.\end{aligned}$$

Fonction de hasard (suite)

On a aussi:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{1}{S(t)}\frac{dS(t)}{dt}$$
$$= -\frac{d\ln S(t)}{dt}$$

Propriétés

- $\lambda : \Re_+ \longrightarrow \Re_+$
- $\int_0^t \lambda(u)du < \infty, \forall t > 0$, mais $\int_0^\infty \lambda(u)du = \infty$
- λ n'est généralement pas monotone

Exemples graphiques de hasard

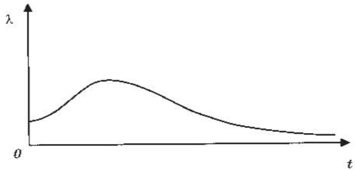


Fig. 12.1. Hazard function for unemployment duration

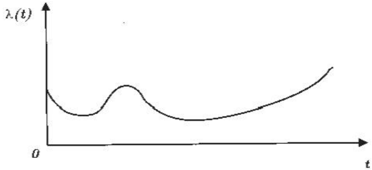


Fig. 12.2. Hazard function for mortality rates among French males

Source: C. Gouriéroux (1989): *Econométrie des variables qualitatives*. Economica

Hazard intégré

$$\Lambda(t) = \int_0^t \lambda(u)du. \tag{8}$$

Propriété:

- $\int_0^t \lambda(u)du < \infty, \forall t > 0$, mais $\int_0^\infty \lambda(u)du = \infty$

Relations entre différentes fonctions

Les fonctions $\Lambda(t), S(t), f(t)$ et $F(t)$ sont en bijection:

$$\Lambda(t) = \int_0^t \lambda(u)du = -\int_0^t \frac{d \ln S(u)}{du} du = -[\ln S(u)]_0^t = -\ln S(t). \quad (9)$$

$$S(t) = \exp[-\Lambda(t)]. \quad (10)$$

$$f(t) = \lambda(t)S(t) = \lambda(t)\exp[-\Lambda(t)]. \quad (11)$$

$$F(t) = 1 - S(t) = 1 - \exp[-\Lambda(t)]. \quad (12)$$

la distribution de T peut-être caractérisée indifféremment une seule de ces fonctions!

Autres relations

Calcul d'espérance:
Si $\lim_{x\rightarrow+\infty} uS(u) \rightarrow 0$ (intégration par partie):

$$\mathrm{E}(T) = \int_0^{\infty} -u\frac{dS(u)}{du}du = \int_0^{\infty} S(u)du. \tag{13}$$

Espérance de la durée de vie restante:

$$r(t) = \mathrm{E}(T - t | T > t). \tag{14}$$

Si $\lim_{x\rightarrow+\infty} uS(u) \rightarrow 0$:

$$r(t) = \frac{1}{S(t)}\int_t^{\infty} S(u)du.$$

Démonstration: exercice!

Autres relations (suite)

Fonction de survie conditionnelle $\forall t > t_0$:

$$S(t+t_0|t_0) = \Pr(T > t+t_0|T > t_0) = \frac{S(t+t_0)}{S(t_0)} = \exp\left(-\int_{t_0}^{t} \lambda(u)du\right). \tag{15}$$

Exemple: la loi exponentielle de paramètre θ ($\theta > 0$)

- fonction de densité: $f(t) = \theta \exp(-\theta t), \forall t > 0$
- fonction de survie:

$$
\begin{aligned}
S(t) &= 1 - F(t) \\
&= 1 - \int_0^t \theta \exp(-\theta u) du \\
&= 1 - [-\exp(-\theta u)]_0^t \\
&= 1 - (-\exp(-\theta t) + 1) \\
&= \exp(-\theta t)
\end{aligned}
$$

- fonction de hasard: $\lambda(t) = f(t)/S(t) = \theta$
- fonction de hasard intégré : $\Lambda(t) = \int_0^t \theta du = \theta t$

Exemple: la loi exponentielle (suite)

- fonction de survie conditionnelle:
  $S(t|t_0) = \exp\left(-\int_{t_0}^{t} \theta du\right) = \exp(-\theta(t-t_0))$
- espérance de la durée de vie restante:

$$\begin{aligned}
r(t) &= \frac{1}{\exp(-\theta t)} \int_t^{\infty} \exp(-\theta u) du \\
&= \frac{1}{\exp(-\theta t)} \left[-\frac{1}{\theta}\exp(-\theta u)\right]_t^{\infty} \\
&= \frac{1}{\exp(-\theta t)} \left(\frac{1}{\theta}\exp(-\theta t)\right) \\
&= \frac{1}{\theta}.
\end{aligned}$$

Loi du hasard intégré

Le hasard est positif, donc le hasard intégré est monotone croissant et on peut écrire:

$$\begin{aligned}\Pr[T > t] &= \Pr\left[\int_0^T \lambda(u)du > \int_0^t \lambda(u)du\right] \\ &= \Pr\left[Z > z\right].\end{aligned}$$

On a également: $\Pr[T > t] = \exp(-\int_0^t \lambda(u)du)$. D'où:

$$\Pr[Z > z] = \exp(-z), \tag{16}$$

et Z suit une loi exponentielle de paramètre 1.

Censures

Chapitre 1

Intuition

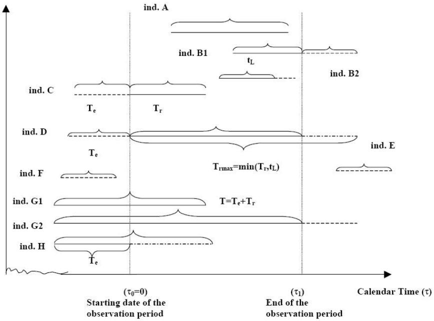
Monsieur i est toujours au chômage à la fin de l'enquête...

Définitions

- Des durées sont généralement censurées, et c'est la raison principale de l'utilisation de la fonction de hasard plutôt que de l'espérance : les hypothèses distributionnelles requises sont plus faibles.
- Soit C une VA positive. Les durées peuvent être censurées à gauche ou à droite:
  - La durée T^* est **censurée à droite** si on observe, au lieu de T^*, le couple (T, δ) où $T = \min(T^*, C)$ et $\delta = 1[T^* < C]$.
  - La durée T^* est **censurée à gauche** si on observe, au lieu de T^*, le couple (T, δ) où $T = \max(T^*, C)$ et $\delta = 1[T^* > C]$.
- les deux censures peuvent être présentes simultanément

Illustration graphique

- A: pas de censure
- B: censure à droite
- C: censure à gauche
- D: censure à gauche et à droite
- E: censure entière à droite
- F: censure entière à gauche



Source : D'Addio, A.C. et Rosholm, M. (2002), *Left censoring in duration data: Theory and Applications*, Working Paper 2002-5, University of Aarhus.

Censures à dates fixes

- La variable de censure C peut être dégénérée (constante), on parle alors de censure fixe.
- on a une censure de **Type I** lorsque la date de censure est fixe et connue à l'avance (ex: tests mécaniques)
- Le modèle Tobit présente une variable réelle sujette à une censure de type I en 0
- on a une censure de **Type II** lorsque la période d'observation s'arrête après un nombre prédéterminé de transitions (ex: enquêtes cliniques)

Echantillon observé

Lorsque T^* est censurée par C, on observe l'échantillon iid $(T, \delta)_{i=1,\dots,n}$.
On a alors deux types d'observations:

- les durées non-censurées, complètes, où $T = T^*$ et $\delta = 1$
- les durées censurées, incomplètes, où $T = C$ et $\delta = 0$

Censure indépendante (non-informative)

- Les approches standards supposent une censure indépendante : les paramètres de la loi de C sont non-informatifs sur le paramètres de la loi de T^*. La censure peut alors être traitée comme exogène.
- les censures de type I et II sont non-informatives
- **Intérêt d'une censure non-informative:**
  - facilite l'identification (à suivre)
  - simplifie l'écriture de la vraisemblance (à suivre)
    Intuition: $\Pr[T = t, \delta = 1] = \Pr[T = t|\delta = 1] \Pr[\delta = 1]$. Lorsque la censure est non-informative, $\Pr[T = t|\delta = 1] = \Pr[T = t]$ et $\Pr[\delta = 1]$ peut être ignorée car non-informatif sur T. La densité des observations se ramène à $\Pr[T = t]$ pour une durée non-censurée et $\Pr[T > t]$ pour une durée censurée.

Conséquences sur l'identification

La censure entraîne une perte d'information. La distribution des observations censurées T permet-elle d'identifier la loi de T^*?
$\Leftrightarrow$ si on connaît la loi des observations, peut-on en déduire une **unique** loi pour T^* (via la survie ou le hasard)?
Réponse: Pas nécessairement.

Exemple de non-identification

Soit une censure à droite de type I à la date t_0.

- on ne peut pas identifier la partie de la distribution qui se trouve à droite de t_0
- donc impossible de savoir s'il y a des modes après t_0, la durée de vie restante après t_0, etc.
- on ne peut identifier que $\Pr(T^* > t_0)$

Dans le cas général:

- si la censure n'est pas exogène, on ne peut pas identifier la loi de T à partir des observations
- sinon, l'identification est possible.

Censure non-informative: une hypothèse raisonnable?

- l'hypothèse de censure exogène n'est pas toujours réaliste
- l'hypothèse est généralement plus fragile dans les modèles à risques concurents
  Exemple: lorsqu'un moteur est soumis à plusieurs risques de pannes, les risques sont généralement dépendant de l'état général du moteur. Les risques sont donc corrélés, et la survenance de l'un d'eux censure les autres.

Vraisemblance en cas de censure exogène

Hypothèses:

- censure à droite exogène ($C \perp\!\!\!\perp T^*$)
- régresseurs constants dans le temps
- la densité de T^* est entièrement définie par un paramètre θ de dimension finie

Vraisemblance en cas de censure exogène (suite)

Contribution à la vraisemblance de i:

- Si non-censure ($\delta_i = 1$): $\Pr(T \in [t, t+dt[, \delta_i = 1, \theta)$
- Si censure ($\delta_i = 0$): $\Pr(T \in [t, t+dt[, \delta_i = 0, \theta)$

D'où:

$$\begin{aligned} L_i(\theta) = & \Pr(T^* \in [t_i, t_i + dt[, C > T; \theta)^{\delta_i} \Pr(C \in [t_i, t_i + dt[, C < T; \theta)^{1-\delta_i} \\ = & [\Pr(T^* \in [t_i, t_i + dt[; \theta) \Pr(C > t_i)]^{\delta_i} \\ & [\Pr(C \in [t_i, t_i + dt[) \Pr(t_i < T; \theta)]^{1-\delta_i} \end{aligned}$$

Vraisemblance en cas de censure exogène (suite)

La censure est indépendante (ne dépend pas de θ). Pour estimer θ, la vraisemblance se réduit à:

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} L_i \\
&= \prod_{i=1}^{n} \Pr(T^* \in [t_i, t_i + dt[; \theta)^{\delta_i} \Pr(T > t_i; \theta)^{1-\delta_i} \\
&= \prod_{i=1}^{n} f(t_i; \theta)^{\delta_i} S(t_i; \theta)^{1-\delta_i}
\end{aligned}
$$

Remarques sur la vraisemblance

- les propriétés de l'estimateur MV s'appliquent: il est asymptotiquement gaussien, sans biais et efficace (suite plus loin).
- se limiter aux observations non-censurées revient à ne retenir que les durées les plus courtes (si censure à droite) donc à sélectionner l'échantillon : l'estimateur est biaisé
- en cas de censure à gauche, la vraisemblance comprend le terme $1 - S(t_i)$

Estimation par maximum de vraisemblance

Chapitre 1

Quelques rappels

Rappels sur le maximum de vraisemblance

Log-vraisemblances

Echantillon iid $(t_i)_{i=1,\dots,n}$, tiré dans la loi P_θ
Distinguer deux cas:

- s'il n'y a pas de censure, la log-vraisemblance s'écrit:

$$\ln L_n(\theta) = \sum_{i=1}^{n} \ln f(t_i;\theta). \tag{17}$$

- s'il y a censure à droite exogène, la log-vraisemblance s'écrit (cf chap 1):

$$\ln L_n(\theta) = \sum_{i=1}^{n} [\delta_i \ln f(t_i;\theta) + (1-\delta_i) \ln S(t_i;\theta)] \tag{18}$$

Conditions de régularité

H1 les VA $(T_i)_{i=1,\ldots,n}$ sont iid, de même loi de densité $f(t;\theta), \theta \in \Theta \subset \Re^p$

H2 l'espace Θ des paramètres est compact

H3 la vraie valeur inconnue θ_0 du vecteur de paramètre est identifiable: $\forall \theta_0 \neq \theta_1, f(t;\theta_0) \neq f(t;\theta_1), \forall t$.

H4 θ_0 appartient à l'intérieur de Θ

H5 la log-vraisemblance $\ln L_n$ est continue en θ

H6 $\mathrm{E}_{\theta_0}[\ln f(T_i;\theta)]$ existe

H7 $\frac{1}{n} L_n(t;\theta) \overset{p.s.}{\rightarrow} \mathrm{E}_{\theta_0}[\ln f(T_i;\theta)], \forall \theta \in \Theta$.

Convergence de l'EMV

Sous ces conditions, la maximisation de la vraisemblance fournit un estimateur convergent de θ_0:

$$\widehat{\theta}_n \in \arg\max \ln L_n(\theta) \overset{n\to\infty}{\underset{}{\overset{p.s.}{\longrightarrow}}} \theta_0 \tag{19}$$

Efficacité asymtptotique et normalité

Sous:

H8 $\ln L_n(\theta)$ est deux fois continûment dérivable dans un voisinage de θ_0

H9 la matrice d'information de Fisher évaluée en θ_0 existe et est inversible.

$$I(\theta_0) = \mathrm{E}_{\theta_0}\left(\frac{\partial \ln L_n(\theta)}{\partial \theta}\frac{\partial \ln L_n(\theta)}{\partial \theta'}\right)$$
$$= -\mathrm{E}_{\theta_0}\left(\frac{\partial^2 \ln L_n(\theta)}{\partial \theta \partial \theta'}\right).$$

$\widehat{\theta}_n$ est de plus efficace et asymptotiquement normal

Quid en cas de censure?

En général, sous les conditions de régularité habituelles, on peut estimer θ_0 en présence de données censurées

Contre-exemple: On considère une population comprenant deux types d'individus tels que la loi de T est un mélange de deux lois. Si toutes les durées d'un groupe d'individus sont censurées, alors les paramètres propres à ce groupe ne sont pas identifiés.

Hétérogénéité non-observée

Chapitre 1

Intuition

Lorsque les individus sont plus hétérogènes que ce que le modèle ne prend en compte, qu'observe-t-on?

Conséquences de l'hétérogénéité non-observée?

- une partie de l'hétérogénéité des agents n'est pas observée par le statisticien (variables manquantes, ...)
- l'hétérogénéité observée (X) et non-observée affectent la loi de T. En présence d'hétérogénéité non-observée (HNO), des individus avec les mêmes valeurs de X peuvent avoir de "vrais" hasards différents!
- on parle également de modèles de mélange

Modélisation

- l'hétérogénéité non-observée (ou omise) est modélisée par une VA v ($v > 0$), indépendante de l'hétérogénéité observée. En $t = 0$, elle a une densité g, de support V.
- modèle linéaire: HNO $\Rightarrow$ biais variable omise
- modèles de durée (non-linéaires): HNO $\Rightarrow$ biais variable omise + biais dans l'estimation de la dépendance au temps écoulé (à suivre)
  Intuition: les "bons" types (v élevés) transitent rapidement. Avec le temps, la population au risque comprend de plus en plus de "mauvais" types (v faibles)
  $\Rightarrow$ la probabilité empirique de sortie décroît dans le temps
  $\Rightarrow$ le hasard observé est décroissant, même sans dépendance négative au temps du "vrai" hasard!

Survie et densité observées (empiriques, agrégées)

$$
\begin{aligned}
S_m(t) &= \int_V S(t|v)g(v)dv \\
&= \mathrm{E}_g[S(t|v)] \\
f_m(t) &= \int_V f(t|v)g(v)dv \\
&= \mathrm{E}_g[f(t|v)]
\end{aligned}
$$

ces fonctions sont calculées avec la distribution de v en $t = 0$

Hasard observé (empirique, agrégé)

$$
\begin{aligned}
\lambda(t) &= -\frac{d}{dt}\ln S_m(t) = \frac{f_m(t)}{S_m(t)} = \frac{1}{S_m(t)}\int_V f(t|v)g(v)dv \\
&= \frac{1}{S_m(t)}\int_V \lambda(t|v)S(t|v)g(v)dv \\
&= \int_V \lambda(t|v)\frac{S(t|v)g(v)}{S_m(t)}dv
\end{aligned}
$$

Or :

$$
\frac{S(t|v)g(v)}{S_m(t)} = \frac{S(t,v)}{S_m(t)} = g(v|t) \tag{20}
$$

$g(v|t)$ est la densité de V à la date t
$\Rightarrow$ la proportion des différents types d'individus dans l'échantillon évolue dans le temps

Hasard agrégé (suite)

$$\begin{aligned}\lambda(t) &= \int_V \lambda(t|v)g(v|t)dv \\ &= \mathrm{E}_{g_t}[\lambda(t|v)] \\ &= \mathrm{E}_g[\lambda(t|v)|T > t]\end{aligned}$$

le hasard agrégé est calculé avec la loi de v à la date t

Biais de la dépendance au temps

Si $\lambda()$ admet des dérivées à l'ordre 1 continues:

$$\begin{aligned}\frac{\partial}{\partial t}\mathrm{E}_{g_t}[\lambda(t|v)] &= \frac{\partial}{\partial t}\int_V \lambda(t|v)g(v|t)dv \\ &= \int_V \left[\frac{\partial}{\partial t}\lambda(t|v)\right] g(v|t)dv + \int_V \lambda(t|v)\frac{\partial}{\partial t}g(v|t)dv \\ &= \mathrm{E}_{g_t}\left[\frac{\partial}{\partial t}\lambda(t|v)\right] + \int_V \lambda(t|v)\left(\frac{\partial}{\partial t}\frac{S(t|v)g(v)}{S_m(t)}\right)dv\end{aligned}$$

Or:

$$\frac{\partial}{\partial t}\frac{S(t,v)}{S_m(t)} = -\frac{f(t|v)g(v)}{S_m(t)} - \frac{S(t|v)g(v)S_m'(t)}{S_m(t)^2}$$

Et:

$$S'_m(t) = -\int_V f(t|v)g(v)dv \tag{21}$$

D'où:

$$\frac{\partial}{\partial t}\frac{S(t,v)}{S_m(t)} = -\frac{f(t|v)g(v)}{S_m(t)} + \frac{S(t|v)g(v)}{S_m(t)^2}\int_V f(t|v)g(v)dv$$

Et:

$$\begin{aligned}\frac{\partial}{\partial t}\mathrm{E}_{g_t}[\lambda(t|v)] =& \mathrm{E}_{g_t}\left[\frac{\partial}{\partial t}\lambda(t|v)\right] - \int_V \lambda(t|v)\frac{f(t|v)g(v)}{S_m(t)}dv \\ &+ \int_V \lambda(t|v)\frac{S(t|v)g(v)}{S_m(t)}\left(\frac{1}{S_m(t)}\int_V f(t|u)g(u)du\right)dv\end{aligned}$$

Ainsi:

$$\begin{aligned}\frac{\partial}{\partial t}\mathrm{E}_{g_t}[\lambda(t|v)] =& \mathrm{E}_{g_t}\left[\frac{\partial}{\partial t}\lambda(t|v)\right] - \int_V \lambda(t|v)\frac{\lambda(t|v)S(t|v)g(v)}{S_m(t)}dv \\ &+ \left[\int_V \lambda(t|v)\frac{S(t|v)g(v)}{S_m(t)}dv\right]\left[\int_V \lambda(t|v)\frac{S(t|v)g(v)}{S_m(t)}dv\right] \\ &= \mathrm{E}_{g_t}\left[\frac{\partial}{\partial t}\lambda(t|v)\right] - \mathrm{E}_{g_t}[\lambda(t|v)^2] + \mathrm{E}_{g_t}[\lambda(t|v)]^2 \\ &= \mathrm{E}_{g_t}\left[\frac{\partial}{\partial t}\lambda(t|v)\right] - \mathrm{Var}_{g_t}[\lambda(t|v)]\end{aligned}$$

Conclusion: en présence d'hétérogénéité non-observée, l'évolution du hasard agrégé est plus faible que l'évolution moyenne des hasards individuels $\Rightarrow$ lorsque le hasard est constant en t mais de variance non-nulle, le hasard empirique décroît en t

Exemple: mover-stayer

La VA v ne prend que la valeur v_1 ou v_2 (avec $v_2 > v_1$ et $v_2 > 1$). On a $\Pr[v = v_1] = \alpha$ et $\Pr[v = v_2] = 1 - \alpha$.
On suppose le hasard constant : $\lambda(t|v_k) = v_k, \forall k = 1, 2$. D'où:

$$S_m(t) = \mathrm{E}_g[S(t|v)] = \alpha \exp(-v_1 t) + (1-\alpha)\exp(-v_2 t)$$
$$f_m(t) = \mathrm{E}_g[f(t|v)] = \alpha v_1 \exp(-v_1 t) + (1-\alpha) v_2 \exp(-v_2 t)$$

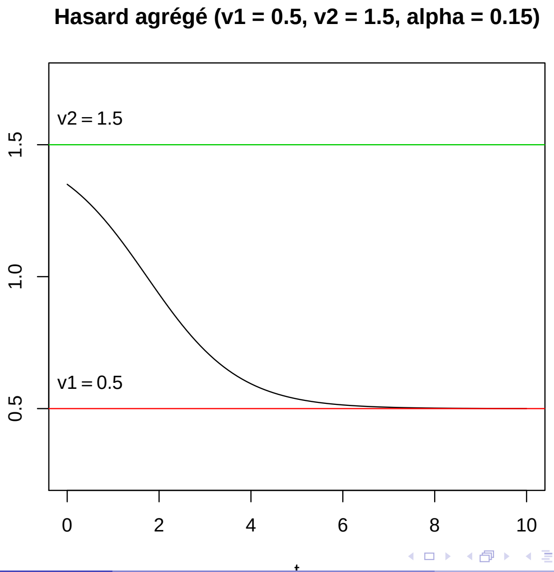
Exemple: mover-stayer (suite)

D'où:

$$\lambda_m(t) = \frac{\alpha v_1 \exp(-v_1 t) + (1-\alpha) v_2 \exp(-v_2 t)}{\alpha \exp(-v_1 t) + (1-\alpha) \exp(-v_2 t)}$$
$$= \frac{\alpha v_1 + (1-\alpha) v_2 \exp(-(v_2 - v_1) t)}{\alpha + (1-\alpha) \exp(-(v_2 - v_1) t)}$$

- si la population était homogène ($v_1 = v_2$), $\lambda_m(t) = v_1$ le hasard empirique serait constant
- comme la population n'est pas homogène ($v_1 \neq v_2$), le hasard empirique est décroissant bien que les hasards individuels soient constants

Exemple: mover-stayer (suite)



Biais sur les paramètres

Hypothèse: On suppose $\lambda(t|x,v) = v\exp(x\beta)$
On peut montrer que (exercice!):

$$\mathrm{E}_g(v|T > t) = -\exp(x\beta)\mathrm{Var}[v|T > t] \leq 0 \tag{22}$$

la valeur moyenne de v décroît au fil du temps parmi les survivant
De plus, on a:

$$\lambda_m(t|x,v) = \mathrm{E}_{g_t}[v\lambda_1(x)] = \exp(x\beta)\mathrm{E}_g[v|T > t] \tag{23}$$

On peut montrer que (exercice!):

$$\frac{\partial}{\partial x}\mathrm{E}_g[v|T > t] = -\exp(x\beta)\beta\mathrm{Var}[v|T > t] \tag{24}$$

Biais sur les paramètres (suite)

Le hasard empirique vérifie donc:

$$\frac{\partial}{\partial x}\ln\lambda_m(t|x,v)=\beta\left[1-\exp(x\beta)\frac{\mathrm{Var}[v|T\geq t]}{\mathrm{E}[v|T\geq t]}\right]<\beta=\frac{\partial}{\partial x}\ln\lambda(t|x,v) \tag{25}$$

Conclusion: la pente du hasard observé est plus faible que la pente du "vrai" hasard

Intuition: les individus ayant des valeurs de x qui augmentent la probabilité de sortie vont sortir rapidement. Après un certain temps, il ne reste plus que les individus avec les valeurs de x qui réduisent la probabilité de sortie. Si on omet v, on sous-estime l'effet de x en lui imputant la variation dans la composition des v.

Echantillonnage

Chapitre 1

Intuition

Comment je construis mon échantillon?

Echantillonnage de stock ou de flux?

But de l'échantillonnage: construire un échantillon représentatif
Deux méthodes de construction d'un échantillon de durées:

- **échantillonnage de stock:** à une date, on interroge des individus dans l'état d'intérêt sur le temps depuis lequel ils y sont
- **échantillonnage de flux:** on enregistre les durées pour les individus qui entrent dans l'état d'intérêt dans un intervalle de temps (on observe les durées depuis le début)

Il est préférable de disposer d'un échantillon dans le flux.
Intuition: les durées longues sont sur-représentées dans l'échantillonnage de stock (et donc les "mauvais" v et x), car elles ont plus de chance de courir à la date d'enquête.

Exemple d'échantillonnage de flux

Défaillances: on ne garde que les entreprises créées après une certaine date
Chômage: on ne garde que les personne entrant dans le chômage après le début de l'enquête

Vraisemblance avec échantillonnage de stock

Supposons que l'individu i soit interrogé alors qu'il est dans l'état actuel depuis un temps s_i^0. Il en sort s_i^1 unités de temps plus tard, d'où $t_i = s_i^0 + s_i^1$.
En l'absence de censure, la vraisemblance s'écrit:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \Pr(T^* \in [t_i, t_i + dt[| T > s_i^0; \theta) \\ &= \prod_{i=1}^{n} \frac{f(t_i;\theta)}{S(s_i^0;\theta)} \\ &= \prod_{i=1}^{n} \lambda(t_i;\theta) \frac{S(t_i;\theta)}{S(s_i^0;\theta)} \end{aligned}$$