Econométrie des modèles de durée

Guillaume Horny*

*Banque de France and UCLouvain

Master 2 MOSEF

Chapitre 3: Modèles paramétriques

Plan

Modèles exponentiel et de Weibull

Chapitre 3

Deux distributions de référence

- **Loi exponentielle:** première loi "naturelle" pour des durées, reliée à la loi de Poisson
- **Loi de Weibull:** une extension de la loi exponentielle, flexible et facile à estimer

Loi de Poisson

Utilisée pour modéliser le **nombre d'événements** par intervalle de temps (nombre d'appels reçus en 1 heure par un standard, nombre de visite chez le docteur par an...)

Inventée par **Siméon Denis Poisson** (1781-1840), publiée en 1838 dans *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*

Notations

Utilise des **processus de Poisson** lorsqu'on s'intéresse à la distribution intertemporelle des événements.
On suppose le temps continu, et on note $N(t)$ le nombre total d'événements qui se sont produit avant t. La suite de VA $\{N(t)|t \geq 0\}$ est un **processus de comptage**. On note également $N(t, t+\Delta), \Delta > 0$ le nombre d'événements se produisant dans l'intervalle $]t, t+\Delta]$.

Définition d'un processus de Poisson

Un processus de Poisson suppose que l'occurrence d'un événement en t est définie par :

$$\begin{aligned}
\Pr\left[N(t, t+\Delta) = 1\right] &= \lambda\Delta + o(\Delta), \\
\Pr\left[N(t, t+\Delta) = 0\right] &= 1 - \lambda\Delta + o(\Delta),
\end{aligned}$$

où $[o(\Delta)/\Delta] \rightarrow 0$ lorsque $\Delta \rightarrow 0$.

Conséquences:

- la probabilité d'occurrence est proportionnelle à longueur de l'intervalle de temps, et le coefficient de proportionnalité est constant (ne varie pas dans le temps)
- l'occurrence est indépendante du nombre d'événements qui se sont déjà produits, et indépendante de t.

De Poisson à l'exponentielle

On peut montrer qu'un processus de Poisson de paramètre λ implique des durées séparant la réalisations de deux événements distribuées selon une loi exponentielle de paramètre λ.

Intuition:

Comme les événements sont indépendants:

$$\Pr[N(t, t + \Delta) = 1] = \Pr\left[N(t, t + \Delta) = 1 | N(0, t)\right]. \tag{1}$$

D'où le hasard:

$$\lim_{\Delta \to 0} \Pr\left[N(t, t + \Delta) = 1 | N(0, t) = 0\right] = \lambda \tag{2}$$

Loi exponentielle

T suit une loi exponentielle de paramètre λ si:

- hasard: $\lambda(t) = \lambda$
- hasard intégré : $\Lambda(t) = \lambda t$
- survie: $S(t) = \exp(-\lambda t)$
- densité: $f(t) = \lambda \exp(-\lambda t), \forall t > 0$

Loi exponentielle (suite)

Propriétés:

- $E(T) = 1/\lambda$ et $V(T) = 1/\lambda^2$
- le hasard intégré est une droite

Introduction de variables explicatives

on écrit: $\lambda_i(t|X_i) = \lambda_i = \exp(X_i\beta)$.

Exemple: code R de MV avec données simulées

```
rm(list=ls())
library(survival)

# simulations des variables explicatives
x1 <- rnorm(1000)
x2 <- rnorm(1000)

# simulation durées exponentielles (S(t) = exp(- lambda t),
# où lambda=exp(X'beta))
lambda0 <- exp(x1 - x2)
t <- (- log(runif(1000))) / lambda0
```

Exemple: code R de MV avec données simulées (suite)

```
# vraisemblance d'un modèle exponentiel
loglikeli <- function(theta){ -1 * sum((theta[1] * x1 \\
+ theta[2] * x2) - t * exp(theta[1] * x1 + theta[2] * x2))}

# maximisation vraisemblance
inits <- c(0,0)

est <- optim(inits, loglikeli, method = "BFGS")
est
```

Exemple: estimation avec STATA

```
streg X, dist(exponential)
```

Loi de Weibull

Inventé par **Waloddi Weibull** (1887-1979), ingénieur et mathématicien suédois. Publia de nombreux travaux sur la résistance des matériaux.

Loi de Weibull

T suit une loi de Weibull de paramètre α et θ si:

- hasard: $\lambda(t) = \alpha\theta t^{\theta-1}$
- hasard intégré : $\Lambda(t) = \alpha t^{\theta}$
- survie: $S(t) = \exp(-\alpha t^{\theta})$
- densité: $f(t) = \alpha\theta t^{\theta-1}\exp(-\alpha t^{\theta})$

Loi de Weibull (suite)

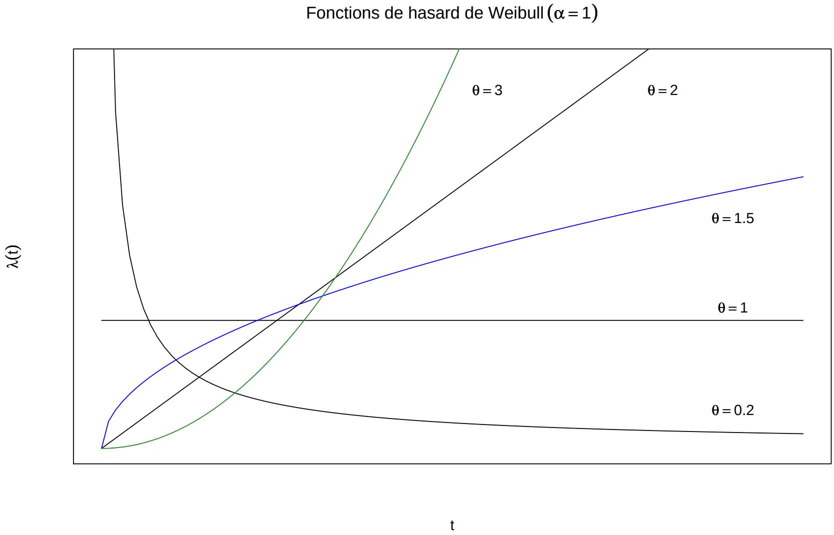
Propriétés:

- pour $\theta = 1$, on retrouve la loi exponentielle
- $d\lambda/dt = \alpha\theta(\theta - 1)t^{\theta-2}$. Le hasard est donc monotone croissant pour $\theta > 1$ et monotone décroissant pour $\theta < 1$
- $\ln \Lambda(t) = \ln \alpha + \ln \theta t$. La courbe $\{\ln t, \ln \Lambda(t)\}$ est une droite

Introduction de variables explicatives

on écrit: $\alpha_i = \exp(X_i\beta)$.

Graphiques de hasards de Weibull



Code R des graphiques

Programme:

```
library(survival)
hasard <- function(x){alpha * theta * x^(theta - 1)}
alpha <- 1
theta <- 1
plot(hasard)
theta <- 2
curve(hasard, add=TRUE)
...
```

Exemple: estimation avec STATA

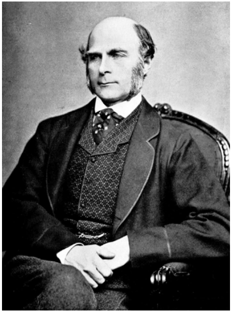
```
streg X, dist(weibull)
```

Modèles log-linéaire et log-logistique

Chapitre 3

Loi log-normale (loi de Galton)

Francis Galton (1822-1911). Cousin de Charles Darwin et explorateur, géographe, météorologue, statisticien, anthropologue, psychométricien, ...
Ses travaux statistiques visent à quantifier les caractéristiques (physiques, psychiques) de l'homme et leur évolution. Auteur de plusieurs contributions à la statistique, dont le modèle linéaire de "régression".
Côté obscur: eugéniste, défendit la "science de l'amélioration des lignées"



Loi log-normale

T suit une loi log-normale de paramètres (μ, σ) si:

- $\ln T \sim \mathcal{N}(\mu, \sigma)$
- densité : $f(t) = \frac{1}{t\sigma} \phi\left(\frac{\ln t - \mu}{\sigma}\right)$
- répartition : $F(t) = \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$
- survie: $S(t) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$
- hasard: $\lambda(t) = \frac{1}{t\sigma} \frac{\phi\left(\frac{\ln t - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)}$

Rappel: changement de variable

Soit $X = g(Z)$, où Z est une variable aléatoire de densité f_Z et g est une application différentiable bijective, alors la densité f_X de X a pour expression:

$$f_X(x) = \frac{f_Z\left(g^{-1}(x)\right)}{\left|g'\left[g^{-1}(x)\right]\right|}. \tag{3}$$

Loi log-normale (suite)

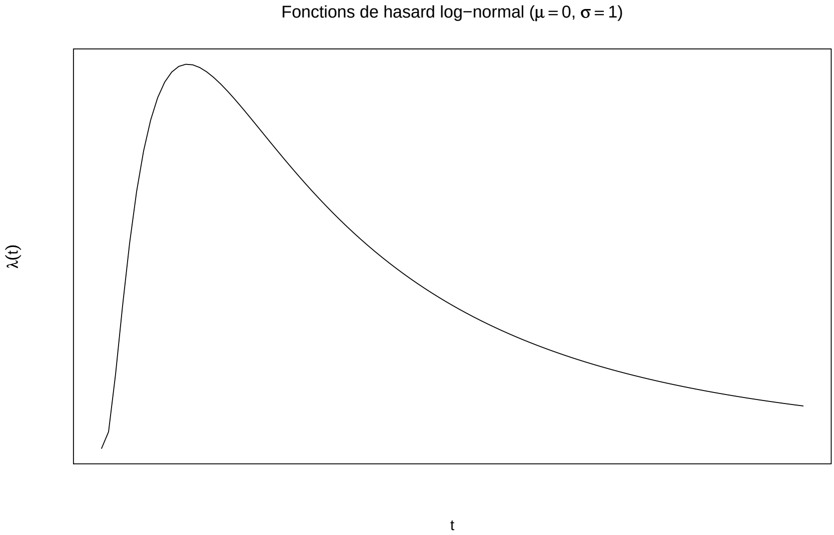
Propriétés:

- $E(T) = \exp(\mu + \sigma^2/2)$ et $Var(T) = [\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2)$.
- $\lim_{t \to 0} \lambda(t) = 0$ et $\lim_{t \to \infty} \lambda(t) = 0$
- la fonction de hasard est croissante puis décroissante, donc unimodale.

Introduction de variables explicatives

on écrit: $\mu_i = X_i \beta$.

Graphique de hasard log-normal



Code R du graphique

Programme:

```
library(survival)
plot(dlnorm)
```

Loi log-logistique

T suit une loi log-logistique de paramètre α et θ si:

- $\ln T \sim \text{Logistique}(\alpha, \theta)$
- densité: $f(t) = \frac{\alpha\theta(\alpha t)^{\theta-1}}{[1+(\alpha t)^{\theta}]^2}$
- répartition: $F(t) = \frac{(\alpha t)^{\theta}}{1+(\alpha t)^{\theta}}$
- survie: $S(t) = \frac{1}{1+(\alpha t)^{\theta}}$
- hasard: $\lambda(t) = \frac{\alpha\theta(\alpha t)^{\theta-1}}{1+(\alpha t)^{\theta}}$

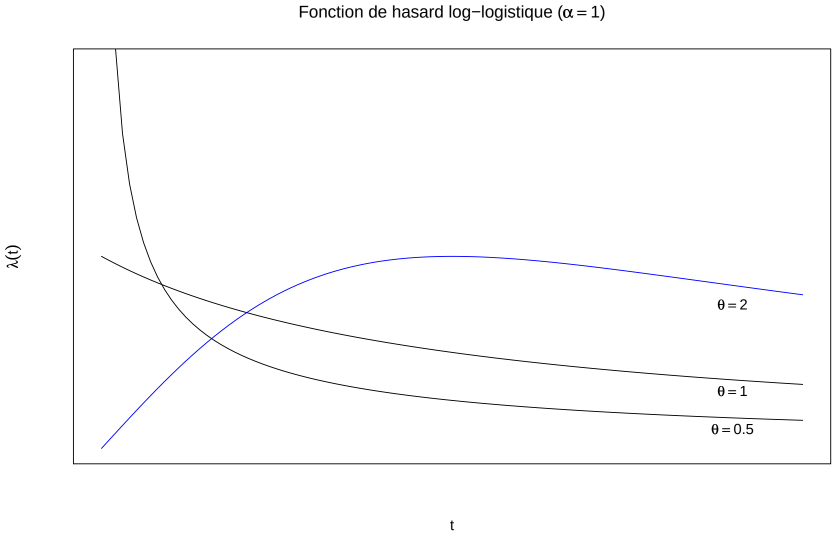
Loi log-logistique (suite)

- pour $\theta < 1$, le hasard est monotone décroissant avec $\lim_{t\to 0} \lambda(t) = \infty$
- pour $\theta = 1$, le hasard est monotone décroissant avec $\lim_{t\to 0} \lambda(t) = \alpha$
- pour $\theta > 1$, le hasard est croissant puis décroissant (donc unimodal), avec $\lim_{t\to 0} \lambda(t) = 0$ et $\lim_{t\to \infty} \lambda(t) = 0$. Le maximum est atteint pour $t = \frac{(\theta-1)^{1/\theta}}{\alpha}$.

Introduction de variables explicatives

on écrit: $\alpha_i = \exp(-X_i\beta)$.

Graphiques de hasards log-logistiques



Code R du graphique

Même principe que pour la loi de Weibull (exercice!)

Exemple: estimation avec STATA

```
streg X, dist(loglogistic)
```

Autres lois

- loi Gamma généralisée
- loi de Gompertz
- loi de Gumbel
- hasards de Box-Cox
- modèle inverse-gaussien
- ...

Exemple: estimation avec STATA

```
foreach model in exponential weibull gompertz lognormal \\
loglogistic gamma {
quietly streg X, dist(‘model’)
estimates store ‘model’
}
estimates stats _all
```

Conclusion

- il existe un grand nombre de lois paramétriques
- sélection de modèles: par tests d'emboîtement, ou par comparaison de critères d'information (AIC,...)

Limitation: modèles peu flexibles (1 ou 2 paramètres, 3 pour la loi Gamma généralisée) pour la dépendance au temps!

Modèles AFT

Chapitre 3

Modèles avec accélération du temps

Les modèles AFT viennent de la modélisation de $\ln T$ plutôt que de T. D'où la régression:

$$\ln t = X'\beta + u, \tag{4}$$

où u a une distribution à support non-nécessairement positif.
$\Rightarrow$ le modèle AFT tire sa flexibilité de la loi de u.

Accélération du temps?

Partant de:

$$\exp(t) = \exp(X'\beta)\exp(u), \tag{5}$$

On peut montrer (exercice!):

$$\lambda(t|X) = \lambda_0[t\exp(-X'\beta)]\exp(X'\beta). \tag{6}$$

Tnterprétation:
Dans un modèle AFT, les X modifient le hasard via un coefficient de proportionnalité appliqué :

- au hasard de base (dilatation verticale),
- à t (dilatation de l'axe des temps, horizontale).

Origine du nom: applications industrielles

Modèles AFT usuels

Distribution de u	Distribution de T
valeur extrême	Weibull
gaussienne	log-normale
logistique	log-logistique

Exemple d'estimation sous R

Code:

```
library(survival)
# ln T ~ Weibull
survreg(Sur(duree, transition) ~ x1 + x2, dist='weibull')
# ln T ~ exponentielle
survreg(Sur(duree, transition) ~ x1 + x2, dist='exponential')
survreg(Sur(duree, transition) ~ x1 + x2, dist='weibull', \\
scale=1 )
# ln T ~ log-normale
survreg(Sur(duree, transition) ~ x1 + x2, dist='lognormal')
```

Exemple d'estimation sous STATA

```
streg X, dist( ) time
```

Interprétation des estimations

$$\ln t = X'\beta + u, \tag{7}$$

- si X_k est en log, β_k est l'élasticité de T par rapport à X_k,
- si X_k est binaire, β_k est approximativement le taux de croissance de T lorsque X_k passe de 0 à 1.

Comparaison HP-AFT

- les coefficients affectent le hasard dans les modèles HP, la durée dans les modèles AFT
- le modèle de Weibull (et donc exponentiel) est à l'intersection des deux classes de modèles
  Exercice: Vérifier en posant u comme distribué selon une loi de valeur extrême.

Remarque générale: vérifier toujours les paramétrisations avant de commenter les résultats!