# Modèle de durée discret avec hétérogénéité non observée

Note de travail

Bernard Lejeune
HEC - Université de Liège

Mars 2014

On s'intéresse au temps qui s'écoule avant qu'un certain évènement se produise. De façon concrète, on considère le temps qui s'écoule avant qu'un chômeur trouve un emploi.

## 1. Variables et notations

On note :

- $T$ = la durée de chômage (mesurée en mois, semaines ou jours) d'un individu pris au hasard dans une population de gens qui débutent un épisode de chômage. Les valeurs possibles de $T$ sont supposées être $t = 0, 1, 2, ...$, $t = 0$ désignant la période d'entrée au chômage, $t = 1$ la première période suivant celle d'entrée au chômage, $t = 2$ la seconde période suivant celle d'entrée au chômage, etc...
- $X$ = un vecteur de variables explicatives (supposées, par commodité, invariantes dans le temps[1]) caractérisant l'individu.
- $V$ = une variable représentant une hétérogénéité non observée. $V$ est supposé indépendant de $X$ : $V \perp X$.

On cherche à modéliser la durée de chômage $T$ d'un individu pris au hasard, étant donné les caractéristiques $X$ de l'individu, et en prenant en compte une possible hétérogénéité non observée $V$.

[1] En pratique, les variables explicatives peuvent être variables dans le temps, pour peu qu'elles soient strictement exogènes.

# 2. Structure du modèle

Un modèle paramétrique est spécifié pour la distribution conditionnelle $T|X,V$ et pour la distribution marginale de l'hétérogénéité non observée $V$ (supposée indépendante de $X$). De ces deux distributions peut être déduite la distribution non conditionnelle à l'hétérogénéité non observée $T|X$.

La distribution conditionnelle $T|X,V$ décrit la façon dont la durée $T$ dépend des variables observées $X$ et non observées $V$. Elle est modélisée au travers d'une fonction de hasard conditionnelle $h(t|X,V)$.

## 2.1. La fonction de hasard conditionnelle $h(t|X,V)$

La fonction de hasard conditionnelle est définie par :

$$\begin{aligned} h(t|X,V) &= I\!P\left[T=t|T\geq t,X,V\right], \quad \forall\, t=0,1,2,... \\ &= g\left[\lambda(t)+\phi(X)+V\right] \end{aligned} \tag{1}$$

La fonction de hasard conditionnelle $h(t|X,V)$ donne la probabilité que la durée $T$ de chômage d'un individu pris au hasard soit (exactement) égale à $t$, sachant qu'il était toujours au chômage en $t-1$ et ses caractéristiques observées $X$ et non observées $V$. La fonction $g\left[.\right]$ peut être spécifiée comme une fonction exponentielle (i.e. $g\left[x\right]=e^x$), ou encore comme une fonction logistique (i.e. $g\left[x\right]=\frac{e^x}{1+e^x}$). Dans le premier cas, le modèle est alors un modèle MPH (Mixed Proportional Hazard) standard. Le terme $\lambda(t)$ est généralement spécifié comme un polynôme ou une fonction constante par intervalle, et le terme $\phi(X)$ comme une combinaison linéaire $X'\gamma$.

## 2.2. La distribution conditionnelle $T|X,V$

La fonction de hasard conditionnelle $h(t|X,V)$ caractérise de façon complète la distribution conditionnelle $T|X,V$. Sa fonction de densité est donnée par[2] :

$$\begin{aligned} f(t|X,V) &= I\!P\left[T=t|X,V\right], \quad \forall\, t=0,1,2,... \\ &= I\!P\left[T=t|T\geq t,X,V\right] I\!P\left[T\geq t|X,V\right] \\ &= h(t|X,V)S(t|X,V) \end{aligned} \tag{2}$$

[2] On note que comme l'événement $(T=t)$ implique l'événement $(T\geq t)$, et donc que l'intersection de ces deux événements est tout simplement l'événement $(T=t)$, on a :

$$I\!P\left[T=t|X,V\right]=I\!P\left[T=t,T\geq t|X,V\right]$$

où $S(t|X,V)$ désigne sa fonction de survie, qui est donnée par :

$$\begin{aligned} S(t|X,V) &= I\!P[T \geq t|X,V] \\ &= \begin{cases} 1 & \text{, si } t=0 \\ \prod_{t^*=0}^{t-1} (1-h(t^*|X,V)) & \text{, } \forall\, t=1,2,... \end{cases} \end{aligned} \tag{3}$$

L'expression (3) de la fonction de survie $S(t|X,V)$ s'obtient par récurrence. Pour $t=0$, on a :

$$S(0|X,V) = I\!P[T \geq 0|X,V] = 1$$

Pour $t=1$, on obtient[3] :

$$\begin{aligned} S(1|X,V) &= I\!P[T \geq 1|X,V] \\ &= I\!P[T \geq 1|T \geq 0,X,V]\, I\!P[T \geq 0|X,V] \\ &= (1 - I\!P[T < 1|T \geq 0,X,V])\, I\!P[T \geq 0|X,V] \\ &= (1 - I\!P[T = 0|T \geq 0,X,V])\, I\!P[T \geq 0|X,V] \\ &= (1-h(0|X,V)) \times 1 = 1 - h(0|X,V) \end{aligned}$$

De même, pour $t=2$, on a[4] :

$$\begin{aligned} S(2|X,V) &= I\!P[T \geq 2|X,V] \\ &= I\!P[T \geq 2|T \geq 1,X,V]\, I\!P[T \geq 1|X,V] \\ &= (1 - I\!P[T < 2|T \geq 1,X,V])\, I\!P[T \geq 1|X,V] \\ &= (1 - I\!P[T = 1|T \geq 1,X,V])\, I\!P[T \geq 1|X,V] \\ &= (1-h(1|X,V))\,(1-h(0|X,V)) \end{aligned}$$

et ainsi de suite.

La fonction de densité conditionnelle $f(t|X,V)$ donne la probabilité que la durée $T$ de chômage d'un individu pris au hasard soit (exactement) égale à $t$, sachant ses caractéristiques observées $X$ et non observées $V$. Pour sa part, la fonction de survie conditionnelle $S(t|X,V)$ donne la probabilité que la durée $T$ de chômage soit supérieure ou égale à $t$, sachant ses caractéristiques observées $X$ et non observées $V$.

[3] On note que comme l'événement $(T \geq 1)$ implique l'événement $(T \geq 0)$, et donc que l'intersection de ces deux événements est tout simplement l'événement $(T \geq 1)$, on a :

$$I\!P[T \geq 1|X,V] = I\!P[T \geq 1, T \geq 0|X,V]$$

[4] Pour les même raisons que ci-dessus, on a :

$$I\!P[T \geq 2|X,V] = I\!P[T \geq 2, T \geq 1|X,V]$$

## 2.3. La distribution de l'hétérogénéité non observée $V$

On spécifie généralement une distribution discrète pour la distribution marginale de $V$, qui par ailleurs est supposée indépendante de $X$. Sa fonction de densité est par définition donnée par :

$$g(v) = I\!P[V = v]$$

L'indépendance de $V$ par rapport à $X$ implique que :

$$I\!P[V = v|X] = I\!P[V = v] = g(v)$$

La fonction de densité $g(v)$ donne la probabilité que la valeur $V$ de l'hétérogénéité non observée d'un individu pris au hasard soit égale à $v$. Elle indique la distribution de l'hétérogénéité non observée dans la population des gens qui débutent un épisode de chômage, i.e., en $t = 0$, dans le flux d'entrée au chômage.

## 2.4. La distribution non conditionnelle à l'hétérogénéité non observée $T|X$

La distribution conditionnelle $T|X, V$ et la distribution distribution marginale de $V$ caractérise finalement la distribution non conditionnelle à l'hétérogénéité non observée $T|X$. Sa fonction de densité est donnée par :

$$\begin{aligned} f(t|X) &= I\!P[T = t|X], \quad \forall\, t = 0, 1, 2, ... \\ &= \sum_{\{v\}} I\!P[T = t|X, V = v]\, I\!P[V = v|X] \\ &= \sum_{\{v\}} I\!P[T = t|X, V = v]\, I\!P[V = v] \\ &= \sum_{\{v\}} f(t|X, v)g(v) = \sum_{\{v\}} h(t|X, v)S(t|X, v)g(v) \end{aligned} \tag{4}$$

La fonction de densité jointe conditionnelle $f(t|X)$ donne la probabilité que la durée $T$ de chômage d'un individu pris au hasard soient (exactement) égale à $t$, sachant ses caractéristiques observées $X$, mais non conditionnellement à l'hétérogénéité non observée $V$.

Généralement, un épisode de chômage n'est observé que sur un nombre fini de périodes. Soit $C$ la dernière période d'observation (= la période de censure à droite). Dans cette situation, la valeur de $T$ ne peut être observée de façon exacte que si $T \leq C$. Lorsque ce n'est pas le cas, il est simplement observé que $T \geq C + 1$. La probabilité conditionnelle à $X$, mais non conditionnelles à $V$, que $T \geq c + 1$ est

donnée par :

$$\begin{aligned} I\!P[T \geq c+1|X] &= \sum_{\{v\}} I\!P[T \geq c+1|X, V=v]\, I\!P[V=v|X]\,, \quad \forall\, c=0,1,2,... \\ &= \sum_{\{v\}} I\!P[T \geq c+1|X, V=v]\, I\!P[V=v] \end{aligned}$$

On sait que :

$$I\!P[T \geq c+1|X, V=v] = S(c+1|X, v)$$

de sorte qu'on a :

$$I\!P[T_u \geq c+1|X] = \sum_{\{v\}} S(c+1|X, v) g(v), \quad \forall\, c=0,1,2,... \tag{5}$$

# 3. Estimation du modèle

On cherche à estimer la fonction de hasard conditionnelle $h(t|X,V)$ d'un individu pris au hasard parmi les individus d'une population cible (les individus qui s'inscrivent comme demandeur d'emploi au cours d'une période donnée : échantillonnage en flux).

On commence par spécifier une forme fonctionnelle $h(t|X,V;\beta)$, qui dépend d'un vecteur de paramètres $\beta \in \Theta_\beta$, pour la fonction de hasard conditionnelle $h(t|X,V)$, forme fonctionnelle que l'on suppose correctement spécifiée, i.e. telle que :

$$h(t|X,V;\beta^o) = h(t|X,V), \text{ pour une valeur } \beta^o \in \Theta_\beta$$

A la fonction de hasard conditionnelle $h(t|X,V;\beta)$ correspond la fonction de survie :

$$S(t|X,V;\beta) = \begin{cases} 1 & , \text{si } t=0 \\ \prod_{t^*=0}^{t-1} (1 - h(t^*|X,V;\beta)) & , \forall\, t=1,2,... \end{cases}$$

qui, si la fonction de hasard est correctement spécifiée, est elle-même correctement spécifiée, i.e. telle que :

$$S(t|X,V;\beta^o) = S(t|X,V), \text{ pour } \beta^o \in \Theta_\beta$$

De même, une forme fonctionnelle $g(v;\alpha)$, qui dépend d'un vecteur de paramètres $\alpha \in \Theta_\alpha$, est spécifiée pour la distribution $g(v)$ de l'hétérogénéité non observée, forme fonctionnelle que l'on suppose à nouveau correctement spécifiée, i.e. telle que :

$$g(v;\alpha^o) = g(v), \text{ pour une valeur } \alpha^o \in \Theta_\alpha$$

Une distribution discrète, entièrement non contrainte, est généralement spécifiée pour $g(v;\alpha)$. Dans ce cas, le vecteur de paramètres $\alpha$ contient, d'une part, les différentes valeurs possibles de $v$, et d'autre part, les probabilités de ces valeurs

possibles.

## 3.1. Echantillonnage

On suppose que l'estimation s'appuye sur un échantillon d'observations de taille $n$ obtenu en tirant au hasard $n$ individus dans la population cible. On note $T_i$ et $X_i$ respectivement la durée de chômage et les caractéristiques observées d'un individu $i$ pris au hasard. On suppose par ailleurs que l'épisode de chômage de chaque individu $i$ est observé durant une période allant de $t = 0$ jusque $t = C_i$, où $C_i$ désigne un point de censure à droite (= la dernière période d'observation)[5].

Etant donné la fenêtre d'observation finie (censure à droite), la durée $T_i$ n'est pas toujours observée de façon exacte. On peut distinguer 2 cas :

- Cas I : $T_i \leq C_i$. Dans ce cas, $T_i$ est observé de façon exacte.
- Cas II : $T_i > C_i$. Dans ce cas, $T_i$ est censuré : il est simplement observé que $T_i \geq C_i + 1$.

Pour ce schéma d'échantillonnage, on peut représenter ce qui est en pratique observé comme une réalisation du triplet $(T_i^*, D_i, X_i)$, où $D_i$ est une variable binaire (auxiliaire) qui prend la valeur 1 si $T_i$ est observé de façon exacte, et 0 sinon, et $T_i^*$ est défini par :

$$T_i^* = \begin{cases} T_i & \text{, si } D_i = 1 \text{ (= Cas I)} \\ C_i + 1 & \text{, si } D_i = 0 \text{ (= Cas II)} \end{cases}$$

i.e., $T_i^*$ est, par définition, égal à la durée effective $T_i$ lorsque celle-ci est observée de façon exacte, et égal à la valeur minimale de cette durée lorsque celle-ci est censurée.

## 3.2. Estimateur du maximum de vraisemblance

L'estimateur maximum de vraisemblance du modèle est défini par :

$$\hat{\theta} = \left(\hat{\beta}', \hat{\alpha}'\right)' = \text{Argmax}_\theta \sum_{i=1}^{n} \ln f^*(T_i^*, D_i | X_i; \theta)$$

où :

$$f^*(t_i^*, d_i | x_i; \theta) = I\!P[T_i^* = t_i^*, D_i = d_i | X_i = x_i; \theta]$$

est la fonction de densité, non conditionnelle à l'hétérogénéité non observée, de $T_i^*, D_i | X_i$.

Lorsque $d_i = 1$ (= Cas I), on a :

$$\begin{aligned} I\!P[T_i^* = t_i^*, D_i = 1 | X_i = x_i; \theta] &= I\!P[T_i = t_i^* | X_i = x_i; \theta] \\ &= \sum_{\{v\}} h(t_i^* | x_i, v; \beta) S(t_i^* | x_i, v; \theta) g(v; \alpha) \end{aligned} \quad (6)$$

où la dernière égalité découle de (4).

[5] Dans la suite, il est implicitement supposé que le mécanisme de censure est 'ignorable', i.e. que $T_i \perp C_i \,| X_i, V_i$ et que $V_i \perp (X_i, C_i)$.

Lorsque $d_i = 0$ (= Cas II), par définition, $t_i^* = c_i + 1$, et on a :

$$\begin{aligned} I\!P[T_i^* = t_i^*, D_i = 0 | X_i = x_i; \theta] &= I\!P[T_i \geq c_i + 1 | X_i = x_i; \theta] \\ &= \sum_{\{v\}} S(c_i + 1 | x_i, v; \beta) g(v; \alpha) \\ &= \sum_{\{v\}} S(t_i^* | x_i, v; \beta) g(v; \alpha) \end{aligned} \tag{7}$$

où l'avant-dernière égalité découle de (5), et la dernière égalité du fait que, par définition, $t_i^* = c_i + 1$.

Etant donné les expressions (6) et (7), l'estimateur maximum de vraisemblance du modèle peut simplement s'écrire :

$$\hat{\theta} = \left(\hat{\beta}', \hat{\alpha}'\right)' = \text{Argmax}_\theta \sum_{i=1}^{n} \ln f^*(T_i^*, D_i | X_i; \theta) \tag{8}$$

où :

$$f^*(T_i^*, D_i | X_i; \theta) = \sum_{\{v\}} \left[h(T_i^* | X_i, v; \beta)\right]^{D_i} S(T_i^* | X_i, v; \beta) g(v; \alpha)$$

Sous des conditions de régularité générales, l'estimateur $\hat{\theta} = \left(\hat{\beta}', \hat{\alpha}'\right)'$ est un estimateur convergent et asymptotiquement normal de $\theta^o = (\beta^{o\prime}, \alpha^{o\prime})'$ :

$$\hat{\theta} \xrightarrow{p} \theta^o \quad \text{et} \quad \sqrt{n}(\hat{\theta} - \theta^o) \xrightarrow{d} \mathcal{N}(0, \Phi_o^{-1})$$

où :

$$\begin{aligned} \Phi_o &= \frac{1}{n} \sum_{i=1}^{n} E\left[\left(\frac{\partial \ln f^*(T_i^*, D_i | X_i; \theta)}{\partial \theta}\right)\left(\frac{\partial \ln f^*(T_i^*, D_i | X_i; \theta)}{\partial \theta}\right)'\right] \\ &= -\frac{1}{n} \sum_{i=1}^{n} E\left[\frac{\partial^2 \ln f^*(T_i^*, D_i | X_i; \theta)}{\partial \theta \partial \theta'}\right] \end{aligned}$$

soit en termes d'approximation utilisable est échantillon fini :

$$\hat{\theta} \approx \mathcal{N}(\theta^o, \hat{\Phi}^{-1}/n)$$

où $\hat{\Phi}$ est un estimateur convergent de $\Phi_o$. On utilise typiquement :

$$\hat{\Phi} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln f^*(T_i^*, D_i | X_i; \hat{\theta})}{\partial \theta \partial \theta'}$$

## 3.3. Episodes multiples

Nous avons jusqu'ici supposé qu'un seul épisode de chômage était observé par individu (single spell). On considère maintenant le cas où, au moins pour une partie de l'échantillon, plusieurs épisodes de chômage sont observés par individu (multiple

spells).

On suppose que la seule dépendance possible entre les différents épisodes de chômage d'un même individu se fait au travers de l'hétérogénéité non observée $V$ : la valeur de l'hétérogénéité non observée $V$ est supposée être la même au cours des différents épisodes de chômage de l'individu. Pour le reste, on suppose que le même modèle et le même schéma d'échantillonnage continue de prévaloir.

Soit $m_i$ le nombre d'épisodes de chômage observés pour l'individu $i$, et $(T^*_{is}, D_{is}, X_{is})$ les variables observées – définies comme ci-dessus à la section échantillonnage – pour l'individu $i$ au cours de son épisode de chômage $s$. Les variables observées pour l'individu $i$ dans l'ensemble de ses épisodes de chômage est noté $(T^*_i, D_i, X_i)$[6].

Sous ces hypothèses, la version 'épisodes multiples' de l'estimateur du maximum de vraisemblance (8) est donné par :

$$\hat{\theta} = \left(\hat{\beta}', \hat{\alpha}'\right)' = \text{Argmax}_\theta \sum_{i=1}^{n} \ln f^*(T^*_i, D_i | X_i; \theta)$$

où :

$$f^*(T^*_i, D_i | X_i; \theta) = \sum_{\{v\}} \left( \prod_{s=1}^{m_i} \left[h(T^*_{is} | X_{is}, v; \beta)\right]^{D_{is}} S(T^*_{is} | X_{is}, v; \beta) \right) g(v; \alpha)$$

Les propriétés asymptotiques de cet estimateur ML sont les mêmes que ci-dessus.

[6] I.e. $T^*_i = (T^*_{i1}, ..., T^*_{im_i})$, $D_i = (D_{i1}, ..., D_{im_i})$, and $X_i = (X_{i1}, ..., X_{im_i})$.