Économétrie 2 : TD11-12

Modèles de Durée

Alice Lapeyre & Claire Leroy

April 13, 2022

Plan de la séance

Modèle de durée

Modélisation de durée dans un état.

- Ex : temps passé au chômage, nombre d'années d'études
- N.B.1: On modélise un seul état
  - e.g. en emploi vs au chômage et non en emploi, au chômage ou inactif
- N.B.2: Une transition pour chaque individu
  - économie du travail: lorsque l'individu retrouve un emploi, il le garde pour toujours.

Deux spécificités :

- Durée = variable aléatoire positive, continue ou discrète
  - N.B. discrète car (i) intrinsèquement discrète (semestres d'études) ou (ii) processus continu observé à intervalle discrets (statut d'emploi observé dans l'enquête emploi)
- Problème de la censure - e.g. distribution de la durée de vie d'une cohorte dont beaucoup de membres sont encore vivants (durée de vie $\geq$ âge actuel)

Modèles de durée - Outils

Durée : T

- durée totale qu'un individu passe dans un certain état de la nature
- variable aléatoire, ≥ 0, de f.d.r. F: $Pr(T \leq t) = F(t), t \geq 0$
- lorsque T est continue, on note f sa densité: f(t)= dF(t)

Fonction de survie : S(t) = Pr(T > t) = 1 - F(t)

- i.e. probabilité que l'individu reste dans cet état au moins t périodes

Fonction de hasard : h(t) $= lim \frac{Pr(T \in]t, t+\eta[| T > t)}{\eta} = \frac{f(t)}{S(t)}$

- i.e. probabilité de quitter cet état en T = t, sachant que l'individu n'a pas quitté l'état avant cela
- NB : $h(t) = -[lnS(t)]'$

Modèles de durée - Outils

Modéliser l'effet de variables X sur $h(t)$

- But : connaître l'effet de certaines caractéristiques X sur la probabilité de sortir d'un certain état en t
- Exemple : $X =$ toucher une allocation chômage ou non. $T =$ durée passée au chômage.
- NB : X peut être fixe ou varier dans le temps (X_t)

Modèles vus en cours :

- Modèle à hasard proportionnel : $h(t|X = x) = \phi(x)h_0(t)$
- Modèle de vie accélérée : $h(t|X = x) = \phi(x)h_0(\phi(x)t)$
  - Que l'on peut écrire aussi : $ln(T) = -ln(\phi(X)) + \varepsilon$
- Modèle à hétérogénéité individuelle : $h(t|\nu, X = x) = \nu\phi(x)h_0(t)$, avec ν l'hétérogénéité individuelle inobservée

Modèles de durée - Estimation

Problème de censure :

- La durée T n'est pas observée pour tout le monde
- On observe $Y = min(T, C)$ et $D = 1(T < C)$
  - D : indicatrice de non-censure
  - Pour les individus non-censurés ($D = 1$) on observe bien $Y = T$
  - Pour les individus censurés ($D = 0$) on observe une durée censurée $Y = C$ (la durée censurée est inférieure à la vraie durée T)

Estimation : cf Ch6 slide 15

- MCO : non convergent, car problème de censure
- Maximum de vraisemblance : convergent
  - Hypothèses paramétriques sur ϕ, h_0 et ν

Plan de la séance

Enoncé

Objet d'étude : durées de chômage T pour des actifs entrant au chômage entre les dates 0 et b > 0

On suppose n'observer que les individus encore au chômage à la date b.
$\rightarrow$ échantillonnage dans un stock $\neq$ échantillonnage dans un flux. (plus de détails dans la question 1)

On note A la date d'entrée au chômage (A $\in$ [0, b]), supposée observée pour les individus de l'échantillon.

Exercice 1 - Question 1

(1) Expliquer intuitivement pourquoi l'échantillon obtenu contient trop de durées de chômage longues.

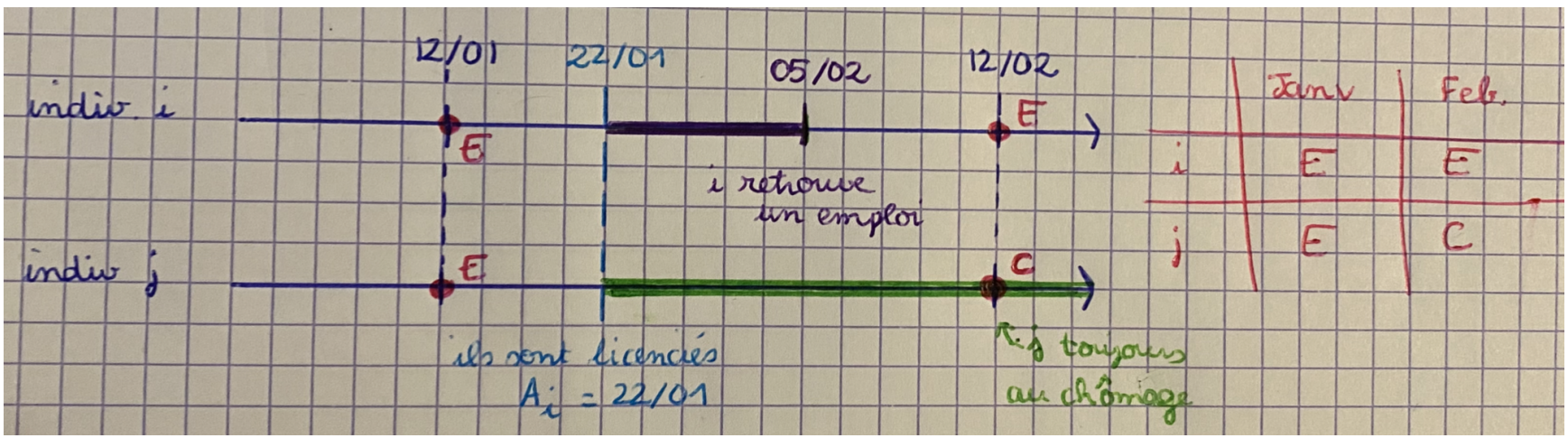
On n'observe que les individus qui sont encore au chômage à la date b. Par conséquent, on n'observe pas les individus qui ont une durée de chômage inférieure à $b - A_i$. Ce faisant, l'échantillon contient trop de durées de chômages longues.

Exercice 1 - Question 1

(1) Expliquer intuitivement pourquoi l'échantillon obtenu contient trop de durées de chômage longues.

Cas pratique: Enquête emploi aux Etats-Unis

- statut d'emploi dans la semaine de référence - celle qui comprend le 12 du mois

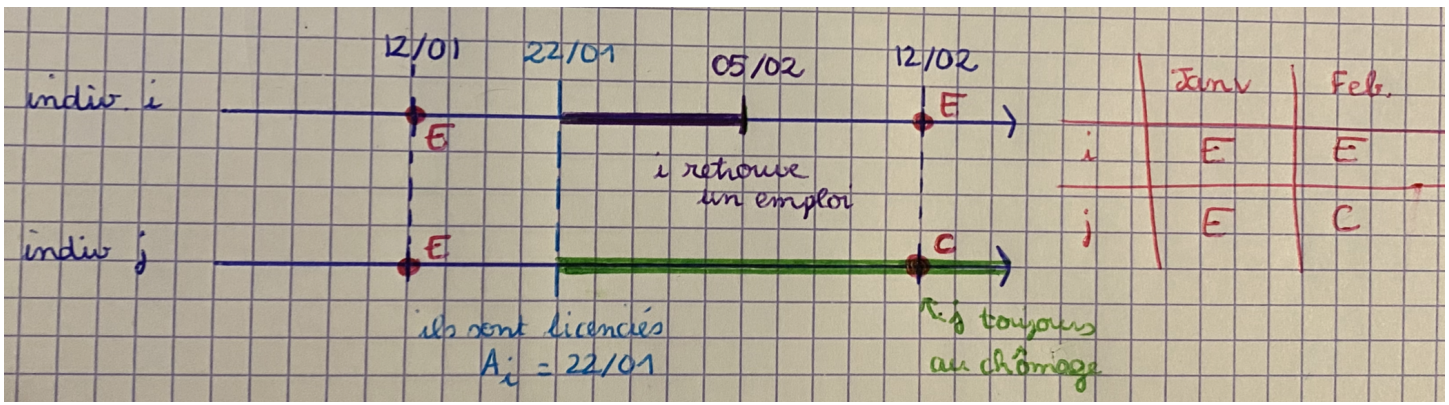


- Parmi les individus tombés au chômage le 22/01 , seuls sont observés ceux qui sont restés au chômage au moins 20 jours (i.e. jusqu'à la date de l'interview). Ceux avec des épisodes de chômage plus courts - comme i - ne font pas parti de notre échantillon de chômeurs.

Exercice 1 - Question 1

(1) Expliquer intuitivement pourquoi l'échantillon obtenu contient trop de durées de chômage longues.

Intuition de la "règle": "on n'observe pas les individus qui ont une durée de chômage inférieure à $b - A_i$"



- Si on normalise t à 0 au 12/01 alors b = 30
- Alors i et j tombent au chômage le 22/01: $A_i = A_j = 10$
- $T_i = 13 < b - A_i$: non observé
- $T_j \geq b - A_j$: observé

Exercice 1 - Question 2

(2) Montrer que la loi de T pour un individu de l'échantillon est non pas la loi marginale de T (i.e. la loi dans la population entière), mais plutôt la loi de $T|S=1$, où l'on définira S en fonction de A et T.

Soit $S_i = 1\{T_i > b - A_i\}$, l'indicatrice d'appartenance à l'échantillon.
La distribution conditionnelle de la durée du chômage dans cet échantillon s'écrit:

$$
\begin{aligned}
F_{T|S=1,X,A}(t) &= \mathbb{P}(T \leq t | S=1, X, A) \\
&= \frac{\mathbb{P}(T \leq t \cap S=1 | X, A)}{\mathbb{P}(S=1|X,A)} \quad \text{car } P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ (Bayes)} \\
&= \frac{\mathbb{P}(b - A < T \leq t | X, A)}{\mathbb{P}(T > b - A | X, A)} \quad \text{par définition de } S_i \\
&= \frac{F_{T|X}(t) - F_{T|X}(b-A)}{S_{T|X}(b-A)} \neq F_{T|X}(t)
\end{aligned}
$$

où la dernière ligne nécessite de supposer $T \perp\!\!\!\perp A|X$ - c'est l'hypothèse à expliciter dans la question suivante

Exercice 1 - Question 2

(2) Montrer que la loi de T pour un individu de l'échantillon est non pas la loi marginale de T (i.e., la loi dans la population entière), mais plutôt la loi de $T|S=1$, où l'on définira S en fonction de A et T .

Distribution conditionnelle de la durée du chômage dans cet échantillon:

$$F_{T|S=1,X,A}(t) = \frac{F_{T|X}(t) - F_{T|X}(b-A)}{S_{T|X}(b-A)}$$

Pour trouver la densité de la durée du chômage dans l'échantillon, il suffit de dériver:

$$\begin{aligned} f_{T|S=1,X,A}(t) &= F'_{T|S=1,X,A}(t) \\ &= \frac{f_{T|X}(t)}{S_{T|X}(b-A)}. \end{aligned}$$

N.B.: on dérive une fonction de la forme: $g(x) = [f(x) - f(a)]/h(a)$. On a donc $g'(x) = f'(x)/h(a)$

Exercice 1 - Question 3

(3) On suppose tout d'abord observer, pour tous les individus de l'échantillon, leur durée de chômage. Montrer, à l'aide d'une hypothèse qu'on explicitera, que la log-vraisemblance conditionnelle de l'échantillon $(T_i, X_i, A_i)_{i=1\dots n}$ s'écrit :

$$\mathcal{L}_n(\theta) = \sum_{i=1}^{n} \ln f_{T|X;\theta}(T_i|X_i) - \ln S_{T|X;\theta}(b - A_i|X_i),$$

où $f_{T|X;\theta}$ (resp. $S_{T|X;\theta}$) désigne la densité (resp. survie) conditionnelle de $T|X$, paramètre par θ.

Densité de la durée de chômage de l'échantillon:

$$\begin{aligned} f_{T|S=1,X,A}(t) &= F'_{T|S=1,X,A}(t) \\ &= \frac{f_{T|X}(t)}{S_{T|X}(b-A)}. \end{aligned}$$

Log-densité:

$$\ln f_{T|S=1,X,A}(t) = \ln f_{T|X}(t) - \ln S_{T|X}(b - A).$$

Exercice 1 - Question 3

(3) (...) Montrer, à l'aide d'une hypothèse qu'on explicitera, que la log-vraisemblance conditionnelle de l'échantillon $(T_i, X_i, A_i)_{i=1...n}$ s'écrit :

$$\mathcal{L}_n(\theta) = \sum_{i=1}^{n} \ln f_{T|X;\theta}(T_i|X_i) - \ln S_{T|X;\theta}(b - A_i|X_i),$$

En faisant l'hypothèse que la distribution de T est caractérisée à l'aide d'un paramètre θ, la contribution de l'individu i à la log-vraisemblance:

$$\ln(f_{T|X;\theta}(T_i|X_i)) - \ln(S_{T|X;\theta}(b - A_i|X_i)).$$

En sommant sur toutes les observations (et en prenant en compte l'indépendance des i), on trouve l'expression de $\mathcal{L}_n(\theta)$ dans l'énoncé.

Exercice 1 - Question 4

(4) (i) Supposons qu'il n'y a pas de covariable et $f_{T|\theta}(t) = \theta \exp(-\theta t)$. Montrer alors que l'estimateur du MV $\hat{\theta}$ vérifie:

$$\frac{1}{\widehat{\theta}} = \frac{1}{n} \sum_{i=1}^{n} T_i + A_i - b.$$

Hypothèse: T suit une loi exponentielle de paramètre $\theta > 0$.
Dans ce cas, on peut montrer que $S_T(t) = \exp(-\theta t)$.

$$\begin{aligned} S_T(t) &= \mathbb{P}(T > t) \\ &= 1 - F(t) \\ &= 1 - (1 - \exp(-\theta t) \quad \text{f.d.r d'une loi exponentielle} \\ &= \exp(-\theta t) \end{aligned}$$

Exercice 1 - Question 4

(4) (i) Montrer alors que $\hat{\theta}_{EMV}$ vérifie: $\frac{1}{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} T_i + A_i - b$.

La log-vraissemblance devient:

$$\begin{aligned}
\mathcal{L}_n(\theta) &= \sum_{i=1}^{n} \ln f_{T|X;\theta}(T_i|X_i) - \ln S_{T|X;\theta}(b - A_i|X_i) \\
&= \sum_{i=1}^{n} \ln(\theta \exp(-\theta t)) - \ln(\exp(-\theta t)) \\
&\text{en remplaçant } f_{T|\theta}(t) \text{ et } S_T(t) \text{ par leurs expressions} \\
&= \sum_{i=1}^{n} \ln(\theta) - \theta T_i + (b - A_i)\theta,
\end{aligned}$$

Exercice 1 - Question 4

(4) (i) Montrer alors que $\hat{\theta}_{EMV}$ vérifie: $\frac{1}{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} T_i + A_i - b$.

Log-vraissemblance:

$$\mathcal{L}_n(\theta) = \sum_{i=1}^{n} \ln(\theta) - \theta T_i + (b - A_i)\theta,$$

Condition du premier ordre:

$$\begin{aligned}
\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} = 0 &\iff \sum_{i=1}^{n} \frac{1}{\hat{\theta}} - T_i + b - A_i = 0 \\
&\iff n\frac{1}{\hat{\theta}} - \sum_{i=1}^{n}(T_i - b + A_i) = 0 \\
&\iff \frac{1}{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} T_i - b + A_i = 0
\end{aligned}$$

De plus, $\forall \theta > 0$, $\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2} = -n\frac{1}{\hat{\theta}^2} < 0 \Rightarrow \hat{\theta}$ maximum global. C'est bien l'EMV.

Exercice 1 - Question 4

(4) (ii) En déduire que $E(T|S=1) = E(T) + b - E(A|S=1)$. Commenter, en lien avec la question 1.

On remarque que l'équation proposée ($\frac{1}{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} T_i + A_i - b.$) est la contrepartie empirique de la condition théorique suivante:

$$\theta^{-1} = \mathbb{E}(T|S=1) + \mathbb{E}(A|S=1) - b,$$

Or, comme $T \sim \mathcal{E}(\theta)$, $\mathbb{E}(T) = \theta^{-1}$. On a donc bien:

$$\mathbb{E}(T|S=1) = \mathbb{E}(T) + b - \mathbb{E}(A|S=1),$$

D'après l'énoncé $\mathbb{E}(A|S=1) < b$ car $A \in [0, b]$, donc $\mathbb{E}(T|S=1) > \mathbb{E}(T)$ ce qui est cohérent avec la remarque faite à la question 1, selon laquelle on observe trop de durées de chômage longues.

Plan de la séance

Enoncé

Objet d'étude : effet de certains facteurs sur la survie au cancer du sein

On observe à la date t_1 :

- *dead*: indicatrice de non-censure
  - $= 0$ si l'individu est toujours vivant (T censurée)
  - $= 1$ si l'individu est décédé (T non censuré)
  - cf question 1
- t : durée (en mois) depuis laquelle le cancer a été détecté
- le vecteur de variables explicatives X qui inclut:
  - *age*: variable continue
  - *smoking*: continue ou indicatrice (?)
  - *dietfat*: nombre moyen de kcalories lipidiques par semaine

Exercice 2 - Question 1

(1) Relier t et dead à la variable d'intérêt T

Soit T_i la durée de survie de l'individu i au cancer du sein une fois qu'il a été détecté ($t = 0$). N.B: il s'agit d'une durée relative par rapport à la détection. Pour rappel, t est la durée depuis laquelle le cancer a été détecté.

Considérons l'individu i observé en t_1 :

- Cas 1: $dead_{i,t_1} = 1$, le patient est décédé en t_1:
  - on observe la durée de survie $T_i = t \rightarrow$ **cas non-censuré**
- Cas 2: $dead_{i,t_1} = 0$, l'individu est toujours vivant.
  - sa durée de survie est au moins égale à t_1 - i.e. $t_1 < T_i \rightarrow$ **cas censuré**

Exercice 2 - Question 2

(2) (i) Commenter la sortie correspondant à la figure 1. On décrira en particulier le nombre de données censurées.

Commande Stata: *stset t, failure(dead)*
on définit la variable d'intérêt Y et l'indicatrice de non censure D

Sortie Stata:

```
      failure event:  dead != 0 & dead < .
. time interval:  (0, t]
exit on or before:  failure
```

- failure event: dead != 0 dead $<$.
  - c'est l'évènement de "sortie" : décès de l'individu dead $==$ 1
  - ici, on passe par la négative - i.e. indicatrice non-nulle et non-manquante
- obs. time interval: (0, t]
  - on observe les individus de la période 0 (détection du cancer) à la période t

Exercice 2 - Question 2

(2) Commenter la sortie correspondant à la figure 1. On décrira en particulier le nombre de données censurées.

```
      80  total observations
(a)    0  exclusions
------------------------------------------------------------
      80  observations remaining, representing
(b)   58  failures in single-record/single-failure data
(c)1257.07  total analysis time at risk and under observation
                                (d)  at risk from t =         0
                     earliest observed entry t =         0
                          last observed exit t =        35
```

- (a): échantillon = 80 obs., toutes inclues dans l'analyse
  - exclusion des durées négatives (aucune ici)
- (b): 58 obs. pour lesquelles la durée est non-censurée
  $\rightarrow$ 80-58 = 22 obs. pour lesquelles ont ne connaît pas la durée de survie exacte mais simplement une borne inf.
- (d) : analyse des failures
  - individus à risque de mourir dès t = 0 (dès le diagnostique)
  - décès le plus tardif en t = 35, i.e. 35 mois après le diagnostique

Exercice 2 - Question 2

(2) (ii) Commenter la sortie correspondant à la figure 1. On décrira en particulier le modèle considéré

Commande Stata: *streg dietfat age smoking, distribution(weibull)*

- On n'a pas besoin de rappeler quelle est la variable expliquée ni la variable de durée
- On renseigne uniquement les variables de contrôle supplémentaires
- Estimation par maximum de vraisemblance

  $\rightarrow$ besoin de spécifier la distribution de $h_0(t)$
- Modèle de Weibull:
  - fonction de survie: $S(t) = \exp(-\lambda t^{\mu})$
  - fonction de hasard: $h(t) = \lambda\mu t^{\mu-1}$
  - taux de hasard décroissant ($\mu < 1$), croissant ($\mu > 1$), ou constant ($\mu = 1$)

Exercice 2 - Question 2

(2) (ii) Commenter la sortie correspondant à la figure 1. On décrira en particulier le modèle considéré

Ici: dans notre modèle de Weibull, on contrôle par un ensemble de variables explicatives ($\lambda = \theta_i := \exp(X_i^T \beta)$). Les fonctions de survie et de hasard s'écrivent donc:

- fonction de survie:
$$S(t) = \exp(\theta_i t^p)$$
- fonction de hasard:
$$h(t) = \theta_i p t^{p-1}$$
- densité:
$$f(t) = \theta_i p t^{p-1} \exp(-\theta_i t^p) 1\{t > 0\}$$

N.B.: Dans le cours la dépendance d'état se note μ. Dans cet exercice et sur les sorties Stata, elle se note p

Exercice 2 - Question 2

(2) (iii) On décrira en particulier l'effet des variables explicatives et la dépendance d'état estimée.

of subjects =	80	Number of obs =	80
of failures =	58		
e at risk =	1257.07		
		LR chi2(3) =	250.96
likelihood =	-13.352142	Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
dietfat	9.22956	2.219297	9.24	0.000	5.761095	14.78621
age	1.749267	.0985256	9.93	0.000	1.566438	1.953436
smoking	5.203393	1.704893	5.03	0.000	2.737728	9.889696
_cons	1.07e-20	4.98e-20	-9.92	0.000	1.22e-24	9.46e-17
/ln_p	1.431728	.0978872	14.63	0.000	1.239872	1.623583
p	4.185925	.4097485			3.455172	5.071228
1/p	.2388958	.0233848			.1971909	.2894212

Ce que l'on va regarder:

- Significativité globale du modèle
- Significativité des coefficients et signe et amplitude de l'effet
- p (μ): dépendance d'état

Exercice 2 - Question 2

(2) (iii) On décrira en particulier l'effet des variables explicatives et la dépendance d'état estimée.

Etape 1: Significativité globale du modèle

```
 of subjects =           80                  Number of obs     =           80
 of failures =           58
e at risk    =      1257.07
                                             LR chi2(3)        =       250.96
 likelihood  =    -13.352142                 Prob > chi2       =       0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
dietfat	9.22956	2.219297	9.24	0.000	5.761095	14.78621
age	1.749267	.0985256	9.93	0.000	1.566438	1.953436
smoking	5.203393	1.704893	5.03	0.000	2.737728	9.889696
_cons	1.07e-20	4.98e-20	-9.92	0.000	1.22e-24	9.46e-17
/ln_p	1.431728	.0978872	14.63	0.000	1.239872	1.623583
p	4.185925	.4097485			3.455172	5.071228
1/p	.2388958	.0233848			.1971909	.2894212

- Estimation par maximum de vraisemblance (pas de R^2)
- Pouvoir explicatif du modèle (p/r à un modèle avec une constante): $LR \sim \chi^2$
- H0: nullité jointe des coefficients
- p-valeur $= 0.000 \rightarrow$ modèle significatif

Exercice 2 - Question 2

(2) (iii) On décrira en particulier l'effet des variables explicatives et la dépendance d'état estimée.

Etape 2: Significativité des coefficients + lecture

Effet d'une variable explicative:

- hazard rate sous forme de ratio: $\beta_0 > 0$ si $exp(\beta_0) > 1$
- `dietfat`: à lire p/r à quelqu'un qui a consommé 0 calories
- `smoking`: à lire p.r. à quelqu'un qui ne fume pas

Exercise 2 - Question 2

(2) (iii) On décrira en particulier l'effet des variables explicatives.

Etape 2: Significativité des coefficients + lecture

```
  of subjects =           80                     Number of obs     =           80
  of failures =           58
ıe at risk    =      1257.07
                                                 LR chi2(3)        =       250.96
ı likelihood  =   -13.352142                     Prob > chi2       =       0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
dietfat	9.22956	2.219297	9.24	0.000	5.761095	14.78621
age	1.749267	.0985256	9.93	0.000	1.566438	1.953436
smoking	5.203393	1.704893	5.03	0.000	2.737728	9.889696
_cons	1.07e-20	4.98e-20	-9.92	0.000	1.22e-24	9.46e-17
/ln_p	1.431728	.0978872	14.63	0.000	1.239872	1.623583
p	4.185925	.4097485			3.455172	5.071228
1/p	.2388958	.0233848			.1971909	.2894212

- Toutes les variables explicatives ont un effet positif et significatif sur la probabilité instantanée de décéder.
- La population fumeuse a une probabilité instantanée de décès qui est égale à 5,20 fois celle de la population non fumeuse.
- Autrement dit, ceteris paribus, fumer par rapport à ne pas fumer multiplie le taux de hasard par 5,20.

Exercice 2 - Question 2

(2) (iiv) On décrira en particulier la dépendance d'état estimée.

Etape 3 : Dépendance d'état p

- p $>$ 1 (resp. $<$ 1) : dépendance au temps positive (resp. négative)
  $\rightarrow$ plus simple de comparer le log p/r à 0

```
of subjects =           80                  Number of obs    =           80
of failures =           58
ıe at risk  =      1257.07
                                            LR chi2(3)       =       250.96
ı likelihood =    -13.352142                Prob > chi2      =       0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
dietfat	9.22956	2.219297	9.24	0.000	5.761095	14.78621
age	1.749267	.0985256	9.93	0.000	1.566438	1.953436
smoking	5.203393	1.704893	5.03	0.000	2.737728	9.889696
_cons	1.07e-20	4.98e-20	-9.92	0.000	1.22e-24	9.46e-17
/ln_p	1.431728	.0978872	14.63	0.000	1.239872	1.623583
p	4.185925	.4097485			3.455172	5.071228
1/p	.2388958	.0233848			.1971909	.2894212

- $p \approx 4.19 > 1$: dépendance d'état (ou dépendance temporelle) positive
- i.e. taux de hasard $h(t)$ croissant avec le temps t
- i.e. temps a un effet positif sur la probabilité de mourir

Exercice 2 - Question 3

(3) (i) La moyenne de t est-elle un estimateur sans biais de $E(T)$?

Non, car il y a problème de censure : on n'observe la durée de survie T que pour ceux qui sont décédés. La moyenne de t sera a priori biaisée à la baisse.

L'espérance théorique de la loi de Weibull est :

$$\mathbb{E}(T_i|X_i) = \theta_i^{-1/p}\Gamma(1+1/p) = \exp(-X_i^T\beta/p)\Gamma(1+1/p)$$

- formule de l'espérance admise : détails ici (page 12)
- on utilise ici une propriété de la fonction exponentielle: $exp(na) = (exp(a))^n$

En intégrant par rapport à X, on obtient :

$$\mathbb{E}(T_i) = \mathbb{E}[\mathbb{E}(T_i|X_i)] = E(\exp(-X_i^T\beta/p))\Gamma(1+1/p)$$

Exercice 2 - Question 3

(3) (ii) Comment peut-on estimer $E(T)$ à partir du modèle ?

En intégrant par rapport à X, on obtient :

$$\mathbb{E}(T_i) = \mathbb{E}[\mathbb{E}(T_i|X_i)] = E(\exp(-X_i^T \beta/p))\Gamma(1+1/p)$$

On pourrait alors prendre sa contrepartie empirique:

$$\widehat{\mathbb{E}(T)} = \Gamma(1+1/\hat{p})\frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{X_i^T \hat{\beta}}{\hat{p}}\right)$$

Exercice 2 - Question 3

(3) (iii) Commenter les sorties ci-dessous.

```
. predict time, mean time

.
end of do-file

. su t
```

Variable	Obs	Mean	Std. Dev.	Min	Max
t	80	15.71337	13.59278	.33	35

```
. display time
48.202061
```

- `predict time, mean time` : prédit la durée de survie moyenne, valeur stockée dans la variable *time*
- *su t* (su = summarize) : stat. desc. de la variable *t* (censurée)
- display time : affiche la valeur prédite - i.e. durée de survie moyenne

Exercice 2 - Question 3

(3) (ii) Commenter les sorties ci-dessous.

```
. predict time, mean time

.
end of do-file

. su t
```

Variable	Obs	Mean	Std. Dev.	Min	Max
t	80	15.71337	13.59278	.33	35

```
. display time
48.202061
```

Conclusion

- Moyenne empirique de t : 15.7
- $\widehat{\mathbb{E}(T)} = 48.2$
- On avait bien un biais vers le bas

Exercice 2 - Question 4

(4) Montrer que l'effet marginal moyen de X_k sur T vérifie : $\Delta_k = -\beta_k E(T)/p$. Comment peut-on estimer Δ_k ?

On a (cf. question précédente) que $\mathbb{E}(T|X) = \exp\left(-\frac{X^T\beta}{p}\right)\Gamma\left(\frac{1}{p}+1\right)$.
Donc, l'effet marginal moyen s'écrit:

$$\begin{aligned}\Delta_k = \mathbb{E}\left[\frac{\partial \mathbb{E}(T|X)}{\partial X_k}\right] &= -\frac{\beta_k}{p} \times \exp\left(-\frac{X^T\beta}{p}\right)\Gamma\left(\frac{1}{p}+1\right) \\ &= -\beta_k \mathbb{E}(T)/p\end{aligned}$$

Ce qui s'estime par:

$$\hat{\Delta}_k = -\widehat{\beta_k}\widehat{\mathbb{E}(T)}/\hat{p}$$

Remarque: l'effet d'une variable sur le taux de défaillance (β_k) est du signe opposé à son effet sur la durée de survie moyenne (Δ_k).

Exercice 2 - Question 5

(5) On considère un modèle avec la constante seule et un modèle incluant seulement age et smoking. Commenter les résultats.

Etape 1: Lecture des coefficients associés aux variables explicatives

	`dietfat`	`age`	`smoking`	`cons`
Modèle 1	9.23***	1.75***	5.20***	1.07e-20***
Modèle 2		1.18***	2.47***	0.00***
Modèle 3				.11***

- age : change peu mais `smoking` : coefficient divisé par deux

Intuition? Biais de variable omise dans modèle 2 potentiellement

- corr(`smoking`,`dietfat`) < 0 : fumeurs ayant par exemple tendance à manger moins gras que les autres (?)
- corr(`died`,`dietfat`) > 0 : fumer augmente le risque de décès du cancer du sein

$\Rightarrow$ coefficient de `smoking` biaisé à la baisse.

Exercice 2 - Question 5

(5) On considère maintenant un modèle avec la constante seule et un modèle incluant seulement age et smoking. Commenter les résultats suivants.

Etape 2: Lecture du paramètre de dépendance d'état

$ln(p)$	1.43***	.36***	-.33***
Controles			
dietfat	✓		
age	✓	✓	
smoking	✓	✓	
constante	✓	✓	✓

- rappel de cours: comparaison p (μ) avec 1, ici on prend le log, on compare donc le log à 0
- Modèle 1 & 2 (modèles avec des contrôles): dépendance d'état positive
- Modèle 3 (constante uniquement): dépendance d'état négative

Exercice 2 - Question 5

(5) Commenter les résultats suivants.

Etape 2: Lecture du paramètre de dépendance d'état

$ln(p)$	1.43***	.36***	-.33***
Contrôles			
dietfat	✓		
age	✓	✓	
smoking	✓	✓	
constante	✓	✓	✓

- Dépendance d'état négative = plus le temps passe et moins il est probable que l'on décède
- En réalité, en l'absence de contrôle, les personnes a risque meurent plus vite et les autres vivent plus longtemps
- Fausse intuition que la probabilité de mourir diminue dans le temps car il ne s'agit pas des mêmes individus (age, conso cigarette, conso alim) $\rightarrow$ biais de variable omise

Exercice 2 - Question 6

(6) (i) On considère enfin deux modèles avec hétérogénéité inobservée, incluant ou non la variable dietfat.

Modèle avec hétérogénéité inobservée: c'est-à-dire que pour chaque individu il existe un facteur multiplicatif ν_i, indépendant de X_i, qui modifie le taux de défaillance. Cela revient à supposer un taux de défaillance de la forme:

$$\tilde{h}_T(t; X, \nu) = \nu p \exp(X^T \beta) t^{p-1}$$

où l'on suppose que $\nu \sim f_\nu(v) = \gamma^\gamma v^{\gamma-1} \exp(-\gamma v)/\Gamma(\gamma) 1\{v > 0\}$.

Exercice 2 - Question 6

(6) (ii) Commenter les résultats.

/ln_p	1.087761	.222261	4.89	0.000	.6521376	1.523385
/ln_the	.3307466	.5250758	0.63	0.529	-.698383	1.359876
p	2.967622	.6595867			1.91964	4.587727
1/p	.3369701	.0748953			.2179729	.520931
theta	1.392007	.7309092			.4973889	3.895711

LR test of theta=0: chibar2(01) = 22.57 Prob >= chibar2 = 0.000

- `LR test of theta=0` : test de la nullité de la variance pour la loi de l'hétérogénéité inobservée ν
  - H0: absence d'hétérogénéité inobservée
  - Modèle sans dietfat: p-valeur: .000 $\rightarrow$ on rejette H0: il y a de l'hétérogénéité inobservée
  - Modèle avec dietfat: p-valeur: 1.000 $\rightarrow$ on ne rejette pas H0
  - Lorsque l'on prend en compte le régime alimentaire des individus, il ne persiste plus d'heterogeneité inobservée
- Par ailleurs, les résultats qualitatifs sont globalement inchangés par rapport aux estimations sans hétérogénéité. (modèle à modèle)

Checklist - TD modèles de durée 2

1. **Bien comprendre l'intuition du problème de censure**
   On n'observe la durée Y non-censurée que pour les individus déjà sortis de l'état $\rightarrow \bar{Y}$ sous-estime la vraie durée $E[T]$ passée dans l'état.
2. **Connaître la définition de la fonction de survie S(t) et de la fonction de hasard h(t)** (similaire à l'idée de taux de mortalité en t)
3. **Savoir interpréter les résultats d'estimation de modèle de durée**
   - Significativité globale du modèle $\Rightarrow$ Likelihood Ratio (LR) test
   - Coefficients associés aux variables X_k : effets de la variable X_k sur $h(t)$. Interprétation qualitative et quantitative + significativité. NB : attention Stata reporte les *hazard ratio*
4. **Comprendre ce qu'est la dépendance d'état**
   - Dépendance d'état : effet de la variable temps sur $h(t)$
   - Dépendance d'état positive : probabilité de sortir de l'état en t augmente avec le temps ($h(t)$ croissante avec t)
   - Savoir interpréter la sortie Stata : $p > 1 \Rightarrow$ dépendance d'état positive
5. **Comprendre pourquoi contrôler pour d'autres facteurs affectant $h(t)$ est important** : biais de variable omise