URCA Hugo Harari-Kermadec

2008-2009 harari@ecogest.ens-cachan.fr

Économétrie 2 : données en strates et modèles de durée

Ce cours est fortement inspiré d'un document de travail de l'INSEE de S. Lollivier,
`http://www.insee.fr/fr/publications-et-services/docs_doc_travail/u9702.pdf` .

Introduction

Nous avons vu que les méthodes usuelles ne sont pas adaptées aux cas où la variable dépendante n'est connue que sous forme discrète. Il n'est alors plus question d'utiliser l'estimateur des moindres carrés ordinaires, sous peine d'introduire des biais dans les estimations. La solution consiste à postuler l'existence d'une variable latente continue, dont une discrétisation à partir d'un ensemble de seuils permet d'obtenir la variable observée. C'est à cette variable latente que l'on applique un modèle linéaire. Deux cas de figure peuvent alors se produire selon la nature du phénomène observé : soit les seuils qui permettent la discrétisation de la variable latente sont connus (données en strates), soit ils ne le sont pas (modèles de durée). Comme on le verra, cette distinction apparemment anodine modifie radicalement la nature du problème et les contraintes liées à l'estimation. La première situation se rencontre par exemple lorsqu'une variable continue n'est observée que sous la forme de tranches (notamment pour des impératifs de collecte, ou afin de limiter des problèmes de non réponse sur la variable continue,...) ou encore dans le cas du modèle Tobit simple. Dans ce dernier modèle, la variable est connue soit en clair si elle est en deçà d'un certain seuil , soit sous forme discrète (dépassement du seuil) dans le cas contraire.
Les modèles de durées usuels appartiennent à la même famille que les précédents. Seule la loi du résidu diffère. Dans les cas précédents, ils étaient généralement supposés normaux voire logistiques alors que dans les modèles de durées, les familles sont plus larges. Cette différence d'approche tient au fait que les paramètres des variables explicatives sont dans la pratique assez peu sensibles au choix des résidus. Mais dans les modèles de durée, **c'est précisément la loi des résidus qui nous intéresse** puisqu'elle détermine la loi du hasard de base, et par conséquent les caractéristiques de la loi de la durée (espérance, existence d'un mode dans les taux de sortie). Il faut donc apporter un soin tout particulier au choix de ce résidu.
Lorsque les observations ne sont pas soumises à des phénomènes de censure, l'estimation de des modèles de durée par les moindres carrés ordinaires est licite, sous réserve que l'on postule un hasard de base log-normal. En présence de censure, la situation est analogue à celle du modèle Tobit simple puisqu'une partie des données est connue exactement et une autre au travers de l'appartenance à un intervalle (une demi-droite en l'occurrence). Seule l'optique change puisque fréquemment cet intervalle est variable avec les individus : toutes les dates de censure ne sont pas identiques. Mais cette situation est en fait peu fréquente pour les variables collectées par questionnaire. On propose en général aux individus un système de tranches dans lequel on l'invite à se placer, de sorte que la variable est toujours connue sous la forme de l'appartenance à un intervalle, dont les limites sont le plus souvent finies.

I Données en strates : une généralisation du modèle Tobit

I.1 Observation d'une variable en tranches

Afin de faciliter la collecte de l'information, par exemple lors d'un entretien, on peut recueillir la variable Y sous une forme qualitative. On demande à l'individu de se placer dans un système de tranches (ou strates, ou classes) préalablement définies, dont les limites $C_1, C_2 \ldots$ sont les mêmes pour tous les individus interrogés.
Lorsque la taille de l'échantillon est grande, la perte d'information par rapport à l'observation d'une variable continue est minime, dès lors que le nombre de tranches est suffisant (6 ou 7, voir Lollivier S. et Verger D.). Ceci tient au fait que l'information fournie par les limites de tranches est riche, surtout si l'on tient compte du fait que les déclarations en clair sont fréquemment arrondies.
Lorsque le nombre de tranches est grand (une vingtaine), la dernière contient en général une faible proportion des observations. Si la taille de l'échantillon est suffisante, l'utilisation des moindres carrés ordinaires sur les centres de

tranches fournit alors des résultats proches de ceux obtenus par l'estimation du maximum de vraisemblance avec résidus normaux. En particulier, la sensibilité à la convention adoptée pour la dernière tranche, peu remplie, influence peu les résultats.

I.2 Modèle mixte

Si la variable expliquée est connue en clair dans un sous échantillon et disponible sous formes de tranches sur le complément, la vraisemblance est composée de deux morceaux, l'un correspondant à la fraction des réponses exactes et l'autre à celle des réponses en tranches. Le premier morceau correspond à un produit de densités, le second à un produit de probabilités. Cette situation se produit par exemple lorsque l'on cherche à interroger les individus sur leurs revenus, mais en restant volontairement discret sur les plus élevés. On demande alors le revenu de façon quantitative en deçà d'un certain seuil C_1, mais seulement une réponse qualitative au delà du seuil (du style "oui, mon revenu dépasse C_1").
On est alors dans la situation du modèle Tobit simple. Les modèles mixtes se rencontrent également lorsque les non-réponses à la question quantitative sont "repêchées" au moyen d'une question en tranches. Dans tous les cas, l'estimation par la méthode du maximum de vraisemblance fournit les valeurs de $\hat{\theta}$ et $\hat{\sigma}^2$ comme précédemment, en utilisant à nouveau la PROC LIFEREG.

I.3 Formalisme général

On considère un échantillon d'individus dont les caractéristiques observables sont notées comme toujours X. On cherche à expliquer une variables Y^* au moyen d'un modèle linéaire :

$$Y^* = X\theta + \varepsilon$$

où ε est un résidu centré de densité f et de fonction de répartition F. Comme dans le modèle Tobit, on distingue plusieurs régimes, ici trois :

$$Y_i = \begin{cases} Y_i^* & \text{si } Y_i^* \notin [C_1; C_3] \\ [C_1; C_2] & \text{si } Y_i^* \in [C_1; C_2] \\ [C_2; C_3] & \text{si } Y_i^* \in [C_2; C_3] \end{cases}$$

– Dans un premier cas, la variable expliqué Y_i^* est observable directement sous forme continue. Comme dans le modèle linéaire simple, la probabilité est alors donnée par la densité :

$$f\left(\frac{Y_i^* - X_i\theta}{\sigma}\right) = f\left(\frac{Y_i - X_i\theta}{\sigma}\right).$$

– Dans les deux cas suivants, on n'observe que l'appartenance à un intervalle $[C_j; C_{j+1}]$. L'une des limite peut être infinie (comme pour Tobit). La probabilité d'être dans l'intervalle est alors :

$$F\left(\frac{C_{j+1} - X_i\theta}{\sigma}\right) - F\left(\frac{C_j - X_i\theta}{\sigma}\right).$$

Pour un échantillon non-mixte, il n'y pas de premier cas, on n'observe que des appartenances à des intervalles.

I.4 Estimation par maximum de vraisemblance

Il reste donc à construire la vraisemblance et à la maximiser. On donne ici l'écriture pour J strates :

$$L(\theta, \sigma) = \prod_{\forall j, Y_i \notin [C_j; C_{j+1}]} f\left(\frac{Y_i - X_i\theta}{\sigma}\right) \prod_{j=1}^{J} \prod_{Y_i^* \in [C_j; C_{j+1}]} F\left(\frac{C_{j+1} - X_i\theta}{\sigma}\right) - F\left(\frac{C_j - X_i\theta}{\sigma}\right).$$

I.5 Implémentation sous SAS

On suit le formalisme utilisé pour Tobit avec la variable LOWER. On construit ainsi deux variables, LOWER et UPPER, donnant les bornes des strates pour chaque individu. L'implémentation de Tobit, (LOWER, Y) vu au cours précédent, se réinterprète comme une strate $[Y, Y]$pour $Y > 0$ et $]-\infty; Y]$ pour $Y = 0$ (et donc LOWER manquante), et c'est sur ce modèle que l'on gère les données mixtes :
PROC LIFEREG DATA=donnees ;
BY sexe ;
MODEL (LOWER, UPPER)=X_1 X_2 X_3 / D=NORMAL **NOLOG** NOINT ;
RUN ;

D= Cette option spécifie la loi des résidus, avec comme possibilités entre autres NORMAL, LOGISTIC, WEIBULL.

NOLOG En fonction de la distribution des résidus, SAS passe tout seul au log les variables LOWER etUPPER. On verra par la suite la raison de ce comportement. Il faut donc forcer SAS à ne pas passer au log, en spécifiant l'option NOLOG.

NOINT On peut comme toujours retirer la constante avec l'option NOINT.

II Modèles de durée

A priori, on pourrait traiter une variable de durée comme n'importe quelle variable aléatoire quantitative continue, à ceci près qu'elle prend nécessairement une valeur réelle positive. Ce n'est pas une caractéristique très discriminante, puisqu'on la retrouve dans d'autres thèmes de l'analyse économique, comme par exemple celle des salaires. La référence habituelle à la loi normale nécessite alors une transformation sur les données, en en prenant par exemple le logarithme. Ainsi, une des lois de base en économétrie des salaires est la loi log-normale, qui revient à faire une hypothèse de normalité sur le log de la variable étudiée. Cette distribution est, on le verra, beaucoup moins centrale en économétrie des durées.
La particularité des données de durées provient du fait qu'elles peuvent s'interpréter facilement comme résultant d'un processus stochastique sous-jacent, c'est à dire d'un cheminement aléatoire qui fait passer un individu entre différents états. Ce processus rend ainsi compte des dates de changements d'état de l'individu (vie et mort, emploi et chômage, être parent d'un enfant ou de deux enfants...). La durée d'un état est alors simplement l'écart entre date de début et date de fin d'un état. Les caractéristiques de ce processus conduisent alors à définir de grandes classes de lois de probabilité pour les durées. De plus, certains outils probabilistes particuliers, comme la fonction de survie ou la fonction de hasard, prendront une place plus déterminante dans l'analyse que l'habituelle densité de probabilité, car ils ont l'avantage de s'interpréter très simplement.

II.1 Outils probabilistes

La variable de durée T présente la particularité de prendre nécessairement des valeurs réelles positives. En plus de la densité f et de la fonction de répartition F, on introduit habituellement deux autres notations :
- La fonction de survie $S(t)$ correspond à la probabilité que la durée soit plus grande que t, soit :

$$S(t) = 1 - F(t).$$

- La fonction de hasard $h(t)$ fournit la probabilité que la durée soit comprise entre t et $t + dt$ sachant qu'elle est plus grande que t :

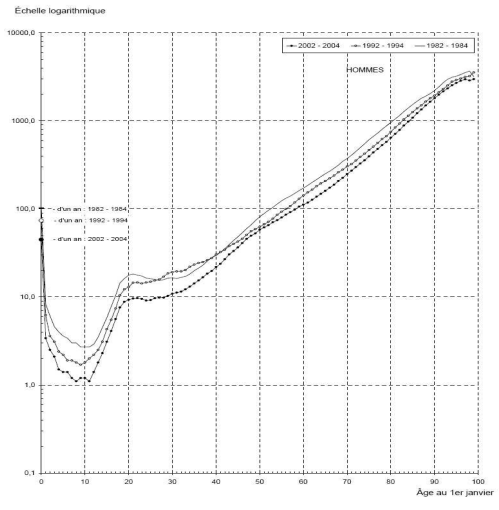
$$h(t) = \frac{f(t)}{S(t)}.$$

  $h(t)$ représente le taux instantané de sortie de l'état que l'on observe. Si l'on s'intéresse par exemple à la durée de vie des individus, il représente le risque de décès à un âge donné sachant que l'on a déjà survécu jusqu'à cet âge. C'est en général cette fonction que chercheront à estimer les modèles économétriques les plus simples. Elle permet de caractériser la probabilité immédiate de changer d'état en t.

On a :

$$h(t) = \frac{\partial \log S(t)}{\partial t} \text{ et } S(t) = \exp\left(-\int_0^t h(u)du\right).$$

Selon les cas étudiés, les fonctions de hasard, ou taux de sortie instantanés, peuvent avoir des formes très différentes. Si l'on considère la durée de vie des hommes en France, le hasard représente simplement le taux de mortalité. Sa forme est en U, avec deux petites « bosses », l'une vers 18-22 ans, l'autre vers 48 ans. La partie décroissante aux tous premiers âges de la vie s'explique par la fin de la période de mortalité néo-natale et infantile, le premier pic par les accidents de la circulation, le second par les maladies cardio-vasculaires. Enfin, le taux de mortalité recommence à augmenter régulièrement aux âges élevés. La représentation d'un tel type de fonction par une loi paramétrée simple n'est, a priori, pas évidente.
Pour d'autres phénomènes étudiés, comme la durée de chômage, cette modélisation peut être plus simple. Ainsi les fonctions de hasard utilisées dans ce cas sont parfois supposées croissantes, puis décroissantes (en raison, par exemple, d'une intensité variable de recherche d'emploi), ou bien simplement décroissantes (en raison, par exemple, d'une réticence des employeurs à embaucher des chômeurs de longue durée).



II.2 Modèles de durée de vie accélérée

Ces modèles font intervenir les variables explicatives, X, comme "accélérateurs" du temps. Ils postulent l'existence d'une loi de référence de la durée T_0. Pour un individu dont les caractéristiques observables sont Z_i, la durée T_i du phénomène étudié s'écrit alors :

$$T_i = T_0 e^{X_i \tau}$$

Tout se passe comme si l'effet des variables observables était d'allonger ou de rétrécir l'unité du temps. L'intérêt principal de ces modèles est en effet de permettre d'interpréter l'effet des variables explicatives comme un changement d'échelle de l'axe du temps. L'égalité précédente conduit à une écriture sous la forme :

$$\log(T_i) = Z_i \tau + \log(T_0).$$

L'espérance de $\log(T_0)$ s'interprète alors comme la constance θ_0. En posant $X_i = (1, Z_i)$ et $\theta = (\theta_0, \tau)'$, on retrouve alors le modèle proche du modèle linéaire :

$$\log(T_i) = X_i \theta + \log(\varepsilon_i).$$

Les résidus ε sont alors supposés d'espérance égale à 1.
Dans le cas très particulier où T_0 suit une loi log-normale, et que toutes les durées sont observables, le modèle de durée peut s'estimer par les moindres carrés ordinaires au moyen de la PROC REG.
Une des particularités les plus fréquentes des modèles de durée est qu'elles sont rarement parfaitement observées. La période d'observation est en effet souvent trop courte pour mesurer les durées les plus longues. On parle alors d'observations censurées. Par exemple, si on suit un échantillon de chômeurs, certains auront quitté cet état à la date de la fin d'observation, d'autres y seront demeurés et la durée totale restera inconnue. En présence de censure, l'ajustement par un modèle linéaire n'est pas envisageable, même si la durée de base suit une loi log-normale.

II.3 Vraisemblance

Si l'on note f et F la densité et la fonction de répartition de $\log(\varepsilon)$, la vraisemblance s'écrit, pour une censure sur la tranche $[C_1; \infty[$:

$$L(\theta) = \prod_{T_i \leq C_1} f(\log(T_i) - X_i\theta) \prod_{T_i > C_1} (1 - F(\log(C_i) - X_i\theta)).$$

Ce modèle est équivalent à celui décrit dans la première partie, avec un seuil C_1 pour les données censurées ? Il est proche du modèle Tobit, mais avec censure à droite. L'estimation ne peut s'opérer qu'au moyen de la proc LIFEREG, en utilisant cette fois (Y,UPPPER).

II.4 LIFEREG pour les données de survie

On applique la procédure LIFEREG comme précédemment, mais sans l'option NOLOG, et avec l'une des distribution suivantes :

option	loi choisie
D=EXPONENTIAL	exponentielle
D=WEIBULL	Weibull
D=LLOGISTIC	log logistique
D=LNORMAL	log normal

La loi exponentielle est une loi de Weibull dont les paramètres sont contraints à 1. SAS propose alors un test (multiplicateur de Lagrange) pour vérifier cette contrainte.