

**INSTITUT NATIONAL DE LA STATISTIQUE ET DES
ETUDES ECONOMIQUES**
Série des Documents de Travail
Méthodologie Statistique

N°9702

MODELES UNIVARIÉS ET MODELES DE DUREE

sur données individuelles

S. LOLLIVIER

Cette note s'inspire pour partie d'un travail réalisé au CREST en collaboration avec C.Casès. Il a également bénéficié des remarques de D.Verger. Toute suggestion est bienvenue en vue d'une version ultérieure.

Ces documents de travail ne reflètent pas la position de l'INSEE et n'engagent que leurs auteurs. Working papers do not reflect the position of INSEE but only their authors views.

RESUME

Ce document est consacré à l'étude des modèles dans lesquels la variable expliquée est soumise à une troncature. Sous un formalisme analogue, ces modèles recouvrent deux grands domaines. Le premier s'intéresse aux modèles pour lequel la variable est observée sous forme d'un système de tranches, le cas échéant sur une partie de l'échantillon, le complément étant observé en clair. Un cas particulier est celui du modèle Tobit simple, dans lequel la variable est observée en clair seulement en dessous d'un certain seuil. On rencontre ce type de situation lorsque l'on a souhaité simplifier la collecte de l'information, par exemple questionner les individus sur une variable sensible sous forme d'un système de tranches plutôt qu'en clair. Le deuxième domaine d'application est celui des modèles de durée. La particularité de ces modèles est de s'intéresser à des variables positives, soumises éventuellement à un phénomène de censure lorsque la durée n'est que partiellement observée.

Dans le texte sont décrits ces différents modèles, leur formalisme, et le moyen de les estimer avec le logiciel SAS. On s'intéresse enfin aux biais introduits par la sélection endogène, qui peuvent apparaître lorsque les durées sont observées à partir de fichiers de stock. C'est par exemple le cas lorsque l'on réalise un suivi des chômeurs à partir d'un échantillon extrait d'une coupe instantanée. Différentes méthodes sont proposées afin d'obtenir des estimateurs convergents des caractéristiques de la loi de la durée.

MOTS CLES : Modèles de durée, modèles qualitatifs, sélection endogène, variables censurées.

Introduction

L'analyse économétrique sur données individuelles cherche généralement à préciser les disparités selon différents critères, mais sur une seule variable. En effet, on s'intéresse principalement à l'étude d'un comportement unique, en négligeant fréquemment d'éventuels phénomènes de simultanéité. Ces derniers se manifestent plutôt lorsque certaines variables explicatives sont suspectées d'endogénéité, auquel cas une instrumentation est nécessaire pour obtenir des estimateurs convergents. Mais on se ramène le plus souvent à un modèle dans lequel une seule variable expliquée apparaît. Un cas particulier très répandu de ce type de modèle est celui pour lequel la variable dépendante est continue (consommation, salaires, patrimoine,...). On utilise alors l'estimateur des moindres carrés ordinaires.

Le problème se complique lorsque 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, soit ils ne le sont pas. 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. Une seule variable intervient contrairement au modèle Tobit généralisé où l'observation de la variable d'intérêt est conditionnée par la valeur d'une autre variable. Les seuils peuvent aussi être inconnus dans toute une série de questions pour lesquels seul un classement est disponible. Ce peut être le cas dans des modèles de choix de portefeuille (détermination du nombre d'actifs patrimoniaux) ou de choix d'intensités (enquêtes d'opinion). La situation extrême est celle des modèles dichotomiques pour lesquelles l'information sur la variable latente est minimale et se réduit à la position par rapport à un seuil inobservé.

Les modèles de durées usuels appartiennent à la même famille que les précédents ; ainsi, les modèles exponentiels, de Weibull, et plus généralement les modèles à durée de vie accélérée s'inscrivent en effet dans une formalisme analogue. 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. Dans les modèles univariés, ces paramètres constituent les variables d'intérêt. Mais dans les modèles de durée, c'est précisément le résidu qui 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 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.

1. Un formalisme général

Un formalisme général régit la plupart de ces modèles univariés, qu'il s'agisse des modèles à variable latente discrétisée ou des modèles de durée à durée de vie accélérée. On considère un échantillon d'individus dont les caractéristiques observables sont notées X . On cherche à expliquer les disparités d'une variable Y au moyen d'un modèle linéaire :

$$Y_i = X_i b + \sigma u_i$$

où u est une variable aléatoire centrée et réduite de densité f et de fonction de répartition F . L'échantillon est constitué de deux sous-populations, dont l'une est éventuellement vide :

↳ Dans un premier sous échantillon E_1 , la variable expliquée Y_i est observable telle quelle sous forme continue (« en clair »). Comme dans le modèle linéaire simple, la probabilité que l'observation appartienne à l'intervalle $[y_i, y_i + dy]$ s'écrit :

$$f\left(\frac{y_i - X_i b}{\sigma}\right)$$

↳ Dans le complément E_2 , seule l'appartenance de la variable à un intervalle $[y_{1i}, y_{2i}]$ est connue. L'une des deux limites de l'intervalle est éventuellement infinie. La probabilité que l'observation appartienne à cet intervalle est alors :

$$F\left(\frac{y_{2i} - X_i b}{\sigma}\right) - F\left(\frac{y_{1i} - X_i b}{\sigma}\right)$$

Au total, la vraisemblance de l'échantillon s'écrit :

$$L = \prod_{i \in E_1} f\left(\frac{y_i - X_i b}{\sigma}\right) \prod_{i \in E_2} \left(F\left(\frac{y_{2i} - X_i b}{\sigma}\right) - F\left(\frac{y_{1i} - X_i b}{\sigma}\right) \right)$$

les fonctions de répartition valant 1 si $y_{2i} = +\infty$ et 0 si $y_{1i} = -\infty$. A ce stade, les limites peuvent être connues ou non. On verra ultérieurement en quoi la résolution diffère selon l'un ou l'autre cas. Tous les modèles paramétriques décrits par la suite se ramèneront à ce formalisme général.

2. Les modèles habituels d'analyse de la variance

2.1. Les moindres carrés ordinaires

Il s'agit d'un cas particulier du modèle précédent pour lesquels la variable expliquée est observée dans tout l'échantillon (E_2 est vide) et la densité est celle d'une loi normale. La maximisation de la vraisemblance conduit à un estimateur \hat{b} qui correspond à celui des moindres carrés ordinaires et à un estimateur $\hat{\sigma}$ asymptotiquement équivalent à celui des moindres carrés ordinaires. C'est ce type de modèle que l'on utilise, ou une forme logarithmique, lorsque les comportements sont observés (salaires, consommation, patrimoine,...). Dans SAS, il s'estime aisément au moyen de la PROC REG ou de la PROC GLM.

2.2. L'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 préalablement définies, dont les limites $\bar{y}_j, j = 1, \dots, J$ sont les mêmes pour tous les individus interrogés. L'ensemble E_1 est alors vide et la vraisemblance du modèle s'écrit :

$$L = \prod_{i \in E} \prod_{j=1}^J \left(F\left(\frac{\bar{y}_{j+1} - X_i b}{\sigma}\right) - F\left(\frac{\bar{y}_j - X_i b}{\sigma}\right) \right)^{I_{y_i \in [\bar{y}_j, \bar{y}_{j+1}]}}$$

avec la convention que $\bar{y}_{j+1} = +\infty$.

Le modèle se présente comme un cas particulier de celui décrit dans la première partie, dans lequel les seuils correspondent aux limites de tranches. Il s'estime par maximisation de la vraisemblance. Tous les paramètres b de même que σ sont identifiables. Les fonctions de répartition les plus couramment utilisées sont celles des lois normales ou logistiques.

Dans SAS, ce type de modèle s'estime facilement au moyen de la PROC LIFEREG, qui, comme son nom ne l'indique pas, permet d'ajuster des modèles sur variables en tranches. La syntaxe de cette procédure permet en effet de définir pour chaque observation une limite basse et une limite haute de tranche, avant d'introduire les variables explicatives et le choix de la loi du résidu.

Ainsi, dans le panel européen des ménages, la variable de patrimoine est en tranches. La PROC LIFEREG permet d'en estimer les disparités selon l'âge (sous forme linéaire par morceaux), le diplôme, le niveau social, la strate de commune, le nombre d'enfants, le type de ménage et le revenu (toutes ces variables sous forme de variables indicatrices). On dispose d'une variable liquid en tranches, à partir de laquelle on crée deux variables de limites de tranches patb et path. Pour les tranches extrêmes, l'une seulement de ces variables est renseignée, l'autre est laissée manquante. Si à la fois patb et path sont manquantes, l'individu n'a pas répondu à la question. Il reste à choisir la loi du résidu et donc le type de modèle (ici log-normal, ce qui correspond au cas le plus fréquent.)

La syntaxe est alors la suivante :

```
Data b;
Set b;
if liquid='01' then do;patb=.;path=20000;end;
else if liquid='02' then do;patb=20000;path=50000;end;
else if liquid='03' then do;patb=50000;path=100000;end;
else if liquid='04' then do;patb=100000;path=300000;end;
else if liquid='05' then do;patb=300000;path=500000;end;
else if liquid='06' then do;patb=500000;path=1000000;end;
else if liquid='07' then do;patb=1000000;path=1500000;end;
else if liquid='08' then do;patb=1500000;path=2000000;end;
else if liquid='09' then do;patb=2000000;path=2500000;end;
else if liquid='10' then do;patb=2500000;path=3000000;end;
else if liquid='11' then do;patb=3000000;path=. ;end;
else
do;patb=. ;path=. ;end;

proc lifereg data=b;weight pondc;
model (patb,path)=vagm30 vag3040 vag4050 vag5060 vag6070 vag70p
diplo0 diplo2-diplo6 nivso0-nivso8 strat0-strat3
nenf1-nenf3 type1-type3 rev1 rev3-rev9 / d=lnormal;
```

ce qui fournit les résultats suivante, interprétables exactement comme ceux d'un modèle d'analyse de variance usuel estimé par la PROC REG.

Lifereg Procedure

Data Set =WORK.B
 Dependent Variable=Log(PATB)
 Dependent Variable=Log(PATH)
 Weight Variable =PONDC
 Noncensored Values= 0 Right Censored Values= 136 ⇨ Description des censures
 Left Censored Values= 819 Interval Censored Values=6013
 Observations with Missing Values= 376

Log Likelihood for LNORMAL -13759.52234 ⇨ Log-Vraisemblance

Lifereg Procedure

| Variable | DF | Estimate | Std Err | ChiSquare | Pr>Chi | Label/Value |
|----------|----|------------|------------------|-----------|--------|------------------------|
| | | \hat{b} | $\hat{\sigma}_b$ | | | |
| INTERCPT | 1 | 9.21003925 | 0.162628 | 3207.245 | 0.0001 | Intercept |
| VAGM30 | 1 | 0.08798331 | 0.014821 | 35.23871 | 0.0001 | |
| VAG3040 | 1 | 0.06704215 | 0.008052 | 69.31701 | 0.0001 | |
| VAG4050 | 1 | 0.02185667 | 0.007934 | 7.589456 | 0.0059 | |
| VAG5060 | 1 | 0.04885419 | 0.008526 | 32.82993 | 0.0001 | |
| VAG6070 | 1 | -0.0065107 | 0.00831 | 0.613822 | 0.4334 | |
| VAG70P | 1 | -0.0283076 | 0.005795 | 23.86381 | 0.0001 | |
| DIPLO0 | 1 | -0.2909669 | 0.052256 | 31.00406 | 0.0001 | |
| DIPLO2 | 1 | 0.06181353 | 0.05658 | 1.193566 | 0.2746 | |
| DIPLO3 | 1 | 0.21581094 | 0.066135 | 10.64849 | 0.0011 | |
| DIPLO4 | 1 | 0.24255722 | 0.084681 | 8.204528 | 0.0042 | |
| DIPLO5 | 1 | 0.26457413 | 0.07575 | 12.19926 | 0.0005 | |
| DIPLO6 | 1 | 0.15801937 | 0.070424 | 5.034757 | 0.0248 | |
| NIVS00 | 1 | 0.46419766 | 0.146861 | 9.990692 | 0.0016 | |
| NIVS01 | 1 | 0.82186552 | 0.089953 | 83.47802 | 0.0001 | |
| NIVS02 | 1 | 0.87957675 | 0.080511 | 119.3553 | 0.0001 | |
| NIVS03 | 1 | 1.05655868 | 0.182963 | 33.34727 | 0.0001 | |
| NIVS04 | 1 | 1.06761488 | 0.16681 | 40.96247 | 0.0001 | |
| NIVS05 | 1 | 0.56812701 | 0.084683 | 45.0086 | 0.0001 | |
| NIVS06 | 1 | 0.39045112 | 0.070652 | 30.54138 | 0.0001 | |
| NIVS07 | 1 | 0.22675185 | 0.067542 | 11.27076 | 0.0008 | |
| NIVS08 | 1 | 0.25768671 | 0.064508 | 15.95732 | 0.0001 | |
| STRAT0 | 1 | 0.55078724 | 0.054838 | 100.8809 | 0.0001 | |
| STRAT1 | 1 | 0.39600277 | 0.05803 | 46.56832 | 0.0001 | |
| STRAT2 | 1 | 0.22431927 | 0.059726 | 14.10597 | 0.0002 | |
| STRAT3 | 1 | 0.05862081 | 0.05033 | 1.356589 | 0.2441 | |
| NENF1 | 1 | 0.00878707 | 0.04706 | 0.034865 | 0.8519 | |
| NENF2 | 1 | 0.07446941 | 0.054417 | 1.872753 | 0.1712 | |
| NENF3 | 1 | -0.1199515 | 0.066119 | 3.291212 | 0.0697 | |
| TYPE1 | 1 | 0.21331312 | 0.127783 | 2.78668 | 0.0951 | |
| TYPE2 | 1 | 0.55880017 | 0.128919 | 18.78807 | 0.0001 | |
| TYPE3 | 1 | 0.49671275 | 0.128143 | 15.02526 | 0.0001 | |
| REV1 | 1 | -0.1939098 | 0.103926 | 3.481356 | 0.0621 | |
| REV3 | 1 | 0.63109612 | 0.067146 | 88.33793 | 0.0001 | |
| REV4 | 1 | 0.98282676 | 0.071125 | 190.9478 | 0.0001 | |
| REV5 | 1 | 1.37415856 | 0.073341 | 351.0562 | 0.0001 | |
| REV6 | 1 | 1.73728459 | 0.075715 | 526.4764 | 0.0001 | |
| REV7 | 1 | 2.25948102 | 0.091496 | 609.8378 | 0.0001 | |
| REV8 | 1 | 2.81499348 | 0.120845 | 542.6264 | 0.0001 | |
| REV9 | 1 | 3.13910012 | 0.21504 | 213.0951 | 0.0001 | |
| SCALE | 1 | 1.2972722 | 0.01246 | | | Normal scale parameter |

⇩ Ecart-type du résidu

Remarques:

↳ lorsque la taille de l'échantillon est grande, la perte d'information par rapport à l'observation d'une variable continue est minimale, 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.

↳ la souplesse de la syntaxe de la procédure permet de réaliser un ajustement même si les tranches sont individualisées, par exemple lorsque l'on demande au ménage un minorant et un majorant de son patrimoine. Les variables `patb` et `path` correspondent alors au minorant et majorant déclarés par le ménage.

2.3. Les modèles "mixtes"

Si la variable expliquée est connue en clair dans un sous échantillon et disponible sous formes de tranches sur le complément, la vraisemblance est, comme dans le cas général, 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 \tilde{y} , mais seulement une réponse qualitative au delà du seuil (du style « oui, mon revenu dépasse \tilde{y} »). La vraisemblance se déduit aisément de la formule générale et s'écrit:

$$L = \prod_{i \in E_1} f\left(\frac{y_i - X_i b}{\sigma}\right) \prod_{i \in E_2} \left(1 - F\left(\frac{\tilde{y} - X_i b}{\sigma}\right)\right)$$

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{b} et $\hat{\sigma}$ comme précédemment, en utilisant à nouveau la PROC LIFEREG. En effet, lorsque la réponse est connue de façon exacte, il suffit d'affecter la même valeur aux deux variables décrivant l'intervalle, cette valeur étant celle déclarée par le ménage. C'est ainsi le cas pour le panel européen des ménages, pour lequel la déclaration de revenu `mnet` est repêchée par une variable en tranches `mesti`.

```

data a;
set a;
rev=input(mnet,6.);
if rev ne . then do;revb=rev;revh=rev;end;
else if mesti='1' then do;revb=.;revh=3000;end;
else if mesti='2' then do;revb=3000;revh=5000;end;
else if mesti='3' then do;revb=5000;revh=7500;end;
else if mesti='4' then do;revb=7500;revh=10000;end;
else if mesti='5' then do;revb=10000;revh=13000;end;
else if mesti='6' then do;revb=13000;revh=20000;end;
else if mesti='7' then do;revb=20000;revh=30000;end;
else if mesti='8' then do;revb=30000;revh=50000;end;
else if mesti='9' then do;revb=60000;revh=. ;end;
else
do;revb=. ;revh=. ;end;

proc lifereg data=a outest=esti;weight pondc;
model (revb,revh)=vagm30 vag3040 vag4050 vag5060 vag6070 vag70p
diplo0 diplo2-diplo6 nivso0-nivso8 strat0-strat3
nenf1-nenf3 type1-type3 / d=lnormal;

```

L'estimation fournit les résultats suivants, en tous points analogues à ceux obtenus pour un ajustement classique pour lequel toutes la variable serait connue pour tous les individus.

Lifereg Procedure

Data Set =WORK.A
 Dependent Variable=Log(REVB)
 Dependent Variable=Log(REVH)
 Weight Variable =PONDC
 Noncensored Values= 5991 Right Censored Values= 16
 Left Censored Values= 41 Interval Censored Values=1100
 Observations with Missing Values= 196

Log Likelihood for LNORMAL -5432.952055

Lifereg Procedure

| Variable | DF | Estimate | Std Err | ChiSquare | Pr>Chi | Label/Value |
|----------|----|------------|----------|-----------|--------|------------------------|
| INTERCPT | 1 | 8.94724012 | 0.049193 | 33080.52 | 0.0001 | Intercept |
| VAGM30 | 1 | 0.03804874 | 0.004578 | 69.08885 | 0.0001 | |
| VAG3040 | 1 | 0.00657975 | 0.002631 | 6.252358 | 0.0124 | |
| VAG4050 | 1 | 0.01301212 | 0.002616 | 24.75031 | 0.0001 | |
| VAG5060 | 1 | 0.00061276 | 0.002806 | 0.047681 | 0.8271 | |
| VAG6070 | 1 | 0.00273921 | 0.002708 | 1.023088 | 0.3118 | |
| VAG70P | 1 | -0.0068922 | 0.001833 | 14.1425 | 0.0002 | |
| DIPLO0 | 1 | -0.0875714 | 0.016836 | 27.05645 | 0.0001 | |
| DIPLO2 | 1 | 0.10708727 | 0.018532 | 33.39155 | 0.0001 | |
| DIPLO3 | 1 | 0.20453983 | 0.021544 | 90.13374 | 0.0001 | |
| DIPLO4 | 1 | 0.23182775 | 0.027617 | 70.46389 | 0.0001 | |
| DIPLO5 | 1 | 0.25263033 | 0.024452 | 106.7421 | 0.0001 | |
| DIPLO6 | 1 | 0.3501983 | 0.022547 | 241.2404 | 0.0001 | |
| NIVSO0 | 1 | -0.2850696 | 0.044571 | 40.90659 | 0.0001 | |
| NIVSO1 | 1 | -0.1791728 | 0.029059 | 38.01637 | 0.0001 | |
| NIVSO2 | 1 | 0.05279525 | 0.026245 | 4.046693 | 0.0443 | |
| NIVSO3 | 1 | 0.64408415 | 0.057812 | 124.1226 | 0.0001 | |
| NIVSO4 | 1 | 0.64762405 | 0.053783 | 144.9979 | 0.0001 | |
| NIVSO5 | 1 | 0.54461179 | 0.026302 | 428.752 | 0.0001 | |
| NIVSO6 | 1 | 0.29294258 | 0.022414 | 170.811 | 0.0001 | |
| NIVSO7 | 1 | 0.1032311 | 0.021524 | 23.0027 | 0.0001 | |
| NIVSO8 | 1 | 0.08905539 | 0.02055 | 18.77956 | 0.0001 | |
| STRAT0 | 1 | -0.1965433 | 0.017722 | 123.0025 | 0.0001 | |
| STRAT1 | 1 | -0.1644878 | 0.018783 | 76.69383 | 0.0001 | |
| STRAT2 | 1 | -0.1569688 | 0.019333 | 65.92187 | 0.0001 | |
| STRAT3 | 1 | -0.1617025 | 0.016225 | 99.33007 | 0.0001 | |
| NENF1 | 1 | 0.1720999 | 0.015399 | 124.9075 | 0.0001 | |
| NENF2 | 1 | 0.1970958 | 0.017823 | 122.288 | 0.0001 | |
| NENF3 | 1 | 0.2458074 | 0.021398 | 131.9591 | 0.0001 | |
| TYPE1 | 1 | -0.4452637 | 0.04063 | 120.1009 | 0.0001 | |
| TYPE2 | 1 | 0.24221608 | 0.041433 | 34.17562 | 0.0001 | |
| TYPE3 | 1 | -0.0047246 | 0.041249 | 0.013119 | 0.9088 | |
| SCALE | 1 | 0.43984377 | 0.003737 | | | Normal scale parameter |

3. Imputations de la variable latente

3.1 Le cas des réponses en tranches

Une fois l'estimation réalisée, on peut tenter de reconstituer la variable latente. En effet, seule une variable quantitative permet de fournir des caractéristiques de moyenne, de moments, de dispersion ou de concentration. La première méthode qui vient à l'esprit consiste à imputer à chaque observation le centre de la tranche (obtenu par une moyenne arithmétique ou géométrique). Cette imputation est par nature sensible à la convention adoptée pour la dernière tranche. Cependant, en présence de beaucoup de tranches, et lorsque la dernière est peu remplie, cette méthode permet de calculer des moyennes de façon acceptable, que ce soit sur l'ensemble ou sur des grosses strates. Néanmoins, les caractéristiques de concentration sont mal approximées. Ceci vient du fait que la variabilité interne aux tranches n'est pas restituée, et donc que celle de la variable ainsi imputée est insuffisante.

On préfère procéder par simulation, c'est à dire par tirage aléatoire dans la loi théorique des résidus. La technique consiste à utiliser la prédiction $X\hat{b}$, puis à tirer des résidus u^* dans la loi théorique jusqu'à ce que la variable simulée $X\hat{b} + \hat{\sigma}u^*$ soit dans la bonne tranche. En cas de non réponse à la question en tranches, on retient le premier tirage. Cette technique donne de bons résultats pour reconstituer l'essentiel de la distribution de la variable latente (voir Lollivier S., Verger D.). Avec les logiciels disponibles, elle est facile à mettre en oeuvre. Elle atteint toutefois ses limites, surtout dans les modèles en logarithme, lorsqu'il s'agit d'imputer des valeurs correspondant à la dernière tranche, qui n'est pas bornée supérieurement. Des tirages de résidus élevés conduisent à des valeurs parfois élevées de la variable simulée. Le caractère atypique de ces tirages est encore renforcé par l'exponentiation. Même si de telles valeurs ne sont pas nécessairement incompatibles avec ce que l'on sait par ailleurs des valeurs extrêmes de la variable, leur imputation à des individus « représentatifs » ultérieurement extrapolés en fonction de leur taux de sondage pose problème. On peut ainsi aboutir à une concentration excessive de la variable, même si les principaux fractiles sont valables. Aucune solution n'est satisfaisante. Une première consiste à borner la dernière tranche avec une information extérieure. Elle évite les tirages trop excessifs, mais ne règle pas le problème de leur représentativité. On pourrait à la limite dupliquer les individus au prorata de leur pondération, mais cette méthode alourdirait fortement l'exploitation. Dans la pratique, un réglage manuel de ces tirages extrêmes est souvent nécessaire.

3.2 Le cas des réponses mixtes (en continu et en tranches)

Dans la mesure où, dans la pratique, la quasi-totalité de l'échantillon fournit une réponse en continu, la tentation est encore plus forte d'imputer une valeur pour les observations comportant une réponse en tranches, ou une non réponse. D'ailleurs, une abondante littérature est disponible sur les techniques d'imputation des non réponses lorsque l'échantillon ne comporte que des réponses en clair (pas de question de rattrapage en tranches). Ces techniques, couramment employées, sont fondées soit sur une modélisation, soit sur l'imputation directe d'une variable déjà observée dans l'échantillon (« hot deck » total ou stratifié), soit sur un mélange des deux (voir N.Caron, document méthodologique).

On utilise ici des méthodes analogues, mais qui prennent en compte le fait que certaines observations sont connues sous la forme d'appartenance à des tranches. L'imputation doit alors respecter cette information supplémentaire. Un premier moyen pour réaliser les imputations consiste à procéder comme dans la section précédente, à savoir simuler un résidu dans la loi conditionnelle, celui-ci étant ajouté à la valeur centrale. On peut également utiliser une seconde méthode, relativement voisine, mais qui s'inspire des procédures de « hot deck » en utilisant davantage l'information contenue dans l'échantillon. En effet, l'ensemble des observations pour lesquelles la variable est connue en clair fournit un ensemble de résidus « observés ». Pour imputer la variable

latente, on ajoute à la valeur centrale un aléa tiré au sort dans cet ensemble, qui respecte l'appartenance de la variable latente à la tranche. Dans la pratique, on commence par mélanger par randomisation les résidus observés et l'on initialise à zéro un compteur de résidus. On considère ensuite les observations où la variable doit être imputée. Si le premier résidu convient pour la première observation, il est retenu, et l'on incrémente de un le compteur. Sinon, on fait défiler le fichier des résidus en incrémentant le compteur, jusqu'à obtenir un résidu adéquat. On passe ensuite à l'observation suivante sans remettre le compteur à zéro. Si les imputations ne portaient que sur des non réponses totales (pas de tranches), cette méthode réalise un tirage avec remise de n résidus parmi N . Elle présente l'avantage de n'introduire qu'une très faible proportion d'aléa tout en reconstituant la distribution de la variable latente. Outre ces propriétés théoriques plus performantes au vu des techniques de redressement, cette méthode réduit fréquemment dans la pratique le nombre de tirages atypiques dans la dernière tranche, ce qui limite le nombre de réglages manuels, toujours insatisfaisants. Elle est néanmoins un peu plus complexe à mettre en oeuvre que la méthode fondée sur les résidus simulés.

4. Les modèles à seuils inconnus

Même s'ils ont été peu évoqués jusqu'ici, la plupart des modèles à réponse qualitative rencontrés dans la littérature théorique et empirique concernent la situation dans laquelle les limites des tranches ne soient pas directement observables. Ce type de modèle est en effet fréquent lorsqu'on recherche les disparités entre individus d'un nombre de biens (durables, financiers...) ou lorsque le questionnaire comporte des questions d'opinion. La variable observée est par nature qualitative, et non quantitative et discrétisée comme dans les cas précédents. En revanche, la variable latente est plus hypothétique et moins « naturelle » que celle obtenue par discrétisation de la variable continue sous jacente. Il s'agit de propension à détenir tel ou tel bien, de penchant à diversifier, ou encore de moral plus ou moins bon.

En toute généralité, le problème est analogue, puisque la vraisemblance a toujours la forme décrite en 2.2. En revanche, et la nuance est de taille, les seuils y_j sont inconnus et doivent par conséquent être estimés dans la procédure. Il se pose alors un problème d'identification ; en effet, les paramètres sont les y_j , b et σ . Ceux-ci n'interviennent dans la vraisemblance que sous la forme y_j/σ et b/σ . De ce fait, toute multiplication des paramètres par un même scalaire conduit à la même valeur de la vraisemblance. On le voit, la non observation des seuils modifie radicalement la nature du problème, alors même que le formalisme est voisin. L'information disponible est appauvrie dans des proportions considérables. Dans SAS, il faut utiliser la PROC LOGISTIC pour estimer ce type de modèle à seuils inconnus (voir le document n°9606 de la collection « Méthodologie Statistique » pour une présentation approfondie de ces modèles).

On contraint en général σ à être égal à 1 pour rendre le modèle identifiable. Mais les paramètres estimés \hat{b} et les seuils estimés \hat{y}_j n'ont plus une interprétation absolue, et fournissent seulement une échelle d'intensité du phénomène expliqué (on a plus ou moins envie de faire telle ou telle chose).

Le cas limite est celui du modèle à un seul seuil. Dans la vraisemblance, ce seuil et la constante interviennent de la même façon. Etant indiscernables l'un de l'autre dans l'estimation, il est nécessaire d'introduire une contrainte pour rendre le modèle identifiable. Dans la pratique, le seuil est fréquemment contraint à zéro. La vraisemblance s'écrit alors :

$$L = \prod_{i \in E} F(X_i c)^{1_{y_i > 0}} (1 - F(X_i c))^{1_{y_i < 0}}$$

en ayant posé $c = -b/\sigma$.

Si F est la fonction cumulative de la loi normale, le modèle est un probit dichotomique. Si F est la fonction cumulative de la loi logistique, il s'agit d'un logit dichotomique.

5. Les 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.

5.1. Rappels de terminologie

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 :

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

↳ 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.

Enfin, la durée moyenne restante est l'espérance de la durée qui reste sachant que l'on a déjà atteint t :

$$r(t) = E(T - t / T > t)$$

C'est par exemple l'espérance de vie à un âge donné, dans l'exemple précédent.

Chacune de ces trois fonctions caractérise la loi d'une variable de durée, au même titre que la densité de probabilité. La plus utilisée est la fonction de hasard. 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 .

Il existe des relations simples entre densité, survie, hasard et durée moyenne restante. Ainsi, on peut aisément montrer que :

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t)$$

$$S(t) = \exp\left(-\int_0^t h(u) du\right)$$

$$E(T) = \int_0^{\infty} S(u) du$$

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 40 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).

5.2. Modèles à durée de vie accélérée

Ces modèles se rattachent directement au formalisme développé dans la première section. Ils postulent l'existence d'une loi de référence de la durée T_0 . Pour un individu dont les caractéristiques observables sont X , la durée s'écrit :

$$T = T_0 \varphi(X, b)$$

La plupart du temps, on choisit la forme :

$$\varphi(X, b) = \exp(Xb)$$

et donc :

$$T = T_0 \exp(Xb)$$

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 :

$$\log(T) = Xb + \log(T_0)$$

où T_0 est une variable aléatoire dont on contraint généralement l'espérance à être égale à 1. De ce fait, la variable aléatoire en logarithme est centrée. Le modèle de durée est alors assimilable à un modèle linéaire. 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. Si l'on note g et G la densité et la fonction de répartition de la variable aléatoire $\log(T_0)$, la vraisemblance s'écrit:

$$L = \prod_{i \in E_1} g(\log t_i - X_i b) \prod_{i \in E_2} (1 - G(\log \hat{t}_i - X_i b))$$

où E_1 est l'échantillon non censuré et E_2 l'échantillon censuré avec \hat{t}_i égal à la durée censurée.

Ce modèle à nouveau est équivalent à celui décrit dans la première partie, avec un seuil \hat{t}_i pour les durées censurées. L'estimation ne peut s'opérer qu'au moyen de la PROC LIFEREG. Lorsque le seuil est le même pour tous, et sous l'hypothèse de hasard log-normal, l'estimation correspond rigoureusement à celle d'un modèle Tobit simple. En général, le seuil est variable selon les individus, et le hasard de base différent. La procédure d'estimation est voisine, et s'effectue aisément, toujours grâce à la PROC LIFEREG.

Dans le cas de la loi log normale, de fonction de répartition Φ , la fonction de survie, la densité et le hasard s'écrivent :

$$S(t) = 1 - \Phi\left(\frac{\log(t) - Xb}{\sigma}\right)$$

$$f(t) = \frac{\Phi\left(\frac{\log(t) - Xb}{\sigma}\right)}{\sigma}$$

$$h(t) = \frac{\Phi\left(\frac{\log(t) - Xb}{\sigma}\right)}{\sigma(1 - \Phi\left(\frac{\log(t) - Xb}{\sigma}\right))}$$

La durée de vie a alors l'espérance d'une loi log normale :

$$E(T / X) = \exp(Xb + \sigma^2/2)$$

Outre la loi log normale, la loi la plus utilisée pour les modèles à durée de vie accélérée est la loi log logistique. Toutes deux permettent de représenter des hasards présentant un mode (fonctions croissantes puis décroissantes). La fonction de survie, la densité et le hasard s'écrivent :

$$S(t) = (1 + t^\alpha \exp(Xb))^{-1}$$

$$f(t) = \frac{\alpha t^{\alpha-1} \exp(Xb)}{(1 + t^\alpha \exp(Xb))^2}$$

$$h(t) = \frac{\alpha t^{\alpha-1} \exp(Xb)}{1 + t^\alpha \exp(Xb)}$$

La fonction de répartition, complément à 1 de la fonction de survie, peut également s'écrire :

$$F(t) = \frac{\exp[\alpha(\log(t) - Xb / \alpha)]}{1 + \exp[\alpha(\log(t) - Xb / \alpha)]}$$

ce qui signifie que la variable $\log(T)$ suit une loi logistique de variance $\pi^2/3\alpha^2$ et d'espérance $-Xb/\alpha$. On est bien dans le cadre des modèles à durée de vie accélérée. Lorsque $\alpha < 1$, le hasard est monotone décroissant de l'infini à zéro. Si $\alpha = 1$, il est monotone décroissant de $\exp(Xb)$ à zéro. Enfin, si $\alpha > 1$, le hasard présente un mode pour

$$t = \frac{(\alpha - 1)^{1/\alpha}}{\exp(Xb)}$$

et est nul en zéro et à l'infini. L'espérance de la durée s'écrit :

$$E(T / X) = \exp(-Xb)B(1 + 1 / \alpha, 1 - 1 / \alpha)$$

où B correspond à la loi BETA. La PROC LIFEREG permet d'estimer ces modèles log logistiques.

5.3. Les modèles à hasard proportionnel

La forme générale du hasard pour ce type de modèle s'écrit :

$$h(t) = \varphi(X, b)h_0(t)$$

$h_0(t)$ est appelé hasard de base. Il correspond au hasard de la population de référence. L'effet des variables explicatives consiste à multiplier par un facteur d'échelle ce hasard de base. Le plus souvent, on adopte la convention :

$$\varphi(X, b) = \exp(Xb)$$

ce qui revient à postuler un facteur d'échelle multiplicatif. Parmi les modèles à hasard proportionnel, on peut citer:

↳ la loi exponentielle pour laquelle le hasard de base est constant. Cela signifie qu'à n'importe quelle date, la probabilité de changer d'état est la même. C'est la raison pour laquelle on dit fréquemment du modèle exponentiel qu'il est « sans mémoire » (le processus sous-jacent est markovien). Ses caractéristiques sont les suivantes:

$$h(t) = \exp(Xb)$$

$$S(t) = \exp[-t \exp(Xb)]$$

$$f(t) = \exp(Xb) \exp[-t \exp(Xb)]$$

et l'espérance de la durée s'écrit :

$$E(T / X) = \exp(-Xb)$$

Notons que la fonction de répartition peut aussi s'écrire :

$$F(t) = 1 - \exp[-\exp(\log(t) + Xb)]$$

de sorte que le modèle exponentiel peut également s'interpréter comme un modèle à durée de vie accéléré du type de ceux décrits dans la partie précédente. Dans SAS, la PROC LIFEREG permet d'estimer les modèles exponentiels.

↳ la loi de Weibull introduit un paramètre α tel que T_0^α suive une loi exponentielle. Les caractéristiques de cette loi sont alors :

$$\begin{aligned} h(t) &= \alpha t^{\alpha-1} \exp(Xb) \\ S(t) &= \exp[-t^\alpha \exp(Xb)] \\ f(t) &= \alpha t^{\alpha-1} \exp(Xb) \exp[-t^\alpha \exp(Xb)] \\ E(T / X) &= \exp(-Xb) \Gamma(1 + 1/\alpha) \end{aligned}$$

où Γ correspond à la loi GAMMA. Le hasard est monotone, croissant si $\alpha > 1$, décroissant si $\alpha < 1$ et constant si $\alpha = 1$. Notons à nouveau que la fonction de répartition peut s'écrire :

$$F(t) = 1 - \exp\left[-\exp\left(\alpha(\log(t) + \frac{Xb}{\alpha})\right)\right]$$

de sorte que le modèle de Weibull peut encore s'interpréter comme un modèle à durée de vie accéléré, que l'on peut estimer au moyen de la PROC LIFEREG.

↳ les modèles à hasard constant par morceaux

Les modèles à hasard constant par morceaux constituent une généralisation du modèle exponentiel, lui apportant davantage de souplesse. En effet, dans ces modèles, le hasard est constant (conditionnellement aux variables explicatives) au cours d'intervalles de durée, dont le nombre et la longueur sont laissés à l'appréciation de l'utilisateur. Plus précisément, le hasard s'écrit

$$h(t) = \begin{cases} \theta_1 & 0 \leq t < 1 \\ \theta_2 & 1 \leq t < 2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \theta_M & M-1 \leq t < \infty \end{cases}$$

avec $\theta_j = \mu_j \exp(Xb)$. On a choisi ici des intervalles de durées égaux (le mois, le trimestre,...). Un formalisme plus général est disponible dans Lancaster [1990]. Dans chacun des intervalles, le modèle est exponentiel conditionnellement aux variables explicatives. L'espérance de la durée est alors :

$$E(Y / X) = \sum_{i=1}^{M-1} \frac{1}{\theta_i} (1 - \exp(-\theta_i)) \exp(-\sum_{j=1}^{i-1} \theta_j) + \frac{1}{\theta_M} \exp(-\sum_{j=1}^{M-1} \theta_j)$$

SAS ne propose pas de procédure d'estimation de ce type de modèles. Ils présentent néanmoins l'avantage de peu contraindre la forme du hasard de base, et de se rapprocher des modèles non paramétriques ou semi-paramétriques (voir infra).

5.4 L'utilisation de la Proc Lifereg pour les modèles de durée

Cette procédure estime des modèles à durée de vie accélérée (ou s'y ramenant) sous la forme :

$$Y = X\beta + \sigma U$$

où $\exp(U)$ suit une loi connue (exponentielle, logistique, normale). Elle fournit en sortie des estimateurs de β et σ . Il reste à en déduire les estimateurs de b et α , introduits comme paramètres dans les modèles décrits précédemment. On les retrouve soit directement, soit par règle de trois, en procédant de la façon suivante :

↪ pour la fonction **log-normale** :

$$F(t) = \Phi\left(\frac{\log(t) - Xb}{\sigma}\right)$$

on a directement les bons estimateurs.

↪ pour la fonction **log-logistique** avec :

$$F(t) = \frac{\exp[\alpha(\log(t) - Xb / \alpha)]}{1 + \exp[\alpha(\log(t) - Xb / \alpha)]}$$

on retrouve les estimateurs requis au moyen de la transformation $\hat{\alpha} = 1 / \hat{\sigma}$ et $\hat{b} = -\hat{\beta} / \hat{\sigma}$.

↪ dans le cas du **modèle exponentiel**,

$$F(t) = 1 - \exp[-\exp(\log(t) + Xb)]$$

on remarque que $\log(T) + Xb = U$, où $\exp(U)$ suit une loi exponentielle d'espérance 1. On obtient $\hat{b} = -\hat{\beta}$, en contraignant $\sigma = 1$.

↪ enfin, dans le cas du **modèle de Weibull**,

$$F(t) = 1 - \exp\left[-\exp\alpha(\log(t) + \frac{Xb}{\alpha})\right]$$

on a de même $\log(T) + \frac{Xb}{\alpha} = \frac{U}{\alpha}$, où $\exp(U)$ suit à nouveau une loi exponentielle d'espérance 1. Par conséquent : $\hat{\alpha} = 1 / \hat{\sigma}$ et $\hat{b} = -\hat{\beta} / \hat{\sigma}$.

5.5 Mise en oeuvre simplifiée (principales options).

```

PROC LIFEREG          < Options 1 > ;
MODEL response = independants < Options 2 > ; } Instructions obligatoires

By variables          ; }
CLASS variables       ; }
OUTPUT                < Options 3 > ; } Instructions facultatives
WEIGHT variables      ; }
  
```

Options 1 :

DATA =

OUTEST = *data* permet de récupérer les estimateurs dans *data*.

COVOUT ajoute la matrice de variance-covariance dans OUTEST.

Options 2 :

* Censor (list) précise l'existence d'une censure à droite (voir LIFETEST supra).

D = précise la distribution

D=EXPONENTIAL modèle exponentiel

D=WEIBULL Weibull

D=LLOGISTIC log logistique

D=LNORMAL log normal

Options 3 :

OUT = *data* précise le nom du *data* de sortie.

Keyword = name avec

CENSORED = variable indicatrice d'une censure

CDF = cumulative

XBETA = $X\beta$.

CLASS joue le même rôle que dans la PROC GLM.

5.4 Exemples d'estimations

Ceux-ci sont tirés du calendrier mensuel de l'enquête annuelle sur l'emploi. On se limite aux femmes ($s = '2'$). On dispose d'une variable de durée de chômage *duration* exprimée en mois ; celle-ci est censurée si la variable *cens* vaut 1. On introduit la classe d'âge *age* comme seule variable explicative. Trois variantes sont proposées, avec le modèle log-normal, le modèle log-logistique et le modèle de Weibull. Dans les trois cas, la lecture directe des coefficients montre que la durée de chômage augmente avec l'âge, dans des proportions assez voisines (si l'on adopte la même normalisation), ce qui montre que l'estimation des coefficients est assez peu sensible au choix du résidu. En revanche, les trois estimations montrent une forme du hasard assez différente, due à une sous représentation des durées courtes dans l'échantillon. Le modèle de Weibull conclue à un hasard un peu décroissant, alors que le modèle log-logistique présente un mode pour les durées courtes. Seul un examen minutieux au moyen des modèles non paramétriques ou semi-paramétriques permet d'en comprendre la raison, à savoir un taux de sortie minoré pour les durées courtes.

La syntaxe des trois estimations est la suivante :

```
proc lifereg data=a(where=(s='2'));  
class age;  
model duree*cens(1)=age / d=lnormal;  
  
proc lifereg data=a(where=(s='2'));  
class age ;  
model duree*cens(1)=age / d=llogistic;  
  
proc lifereg data=a(where=(s='2'));  
class age ;  
model duree*cens(1)=age / d=weibull;
```

Lifereg Procedure
Class Level Information

| Class | Levels | Values |
|-------|--------|-----------------|
| AGE | 8 | 1 2 3 4 5 6 7 8 |

Number of observations used = 1879

Lifereg Procedure

Data Set =WORK.A
 Dependent Variable=Log(DUREE)
 Censoring Variable=CENS
 Censoring Value(s)= 1
 Noncensored Values= 1265 Right Censored Values= 614
 Left Censored Values= 0 Interval Censored Values= 0

Log Likelihood for LNORMAL -2450.416708 ⇒ *Modèle Log Normal*

Lifereg Procedure

| Variable | DF | Estimate | Std Err | ChiSquare | Pr>Chi | Label/Value |
|----------|----|------------|----------|-----------|--------|---------------------------------|
| INTERCPT | 1 | 2.58123764 | 0.150097 | 295.7427 | 0.0001 | Intercept |
| AGE | 7 | | | 56.02357 | 0.0001 | |
| | 1 | -0.7845203 | 0.161513 | 23.59369 | 0.0001 | 1 |
| | 1 | -0.8038823 | 0.162098 | 24.59415 | 0.0001 | 2 |
| | 1 | -0.7045329 | 0.166983 | 17.80166 | 0.0001 | 3 |
| | 1 | -0.6082512 | 0.171072 | 12.64177 | 0.0004 | 4 |
| | 1 | -0.789089 | 0.17224 | 20.98867 | 0.0001 | 5 |
| | 1 | -0.4620755 | 0.19372 | 5.689512 | 0.0171 | 6 |
| | 1 | 0.05443582 | 0.213643 | 0.064922 | 0.7989 | 7 |
| | 0 | 0 | 0 | . | . | 8 ⇒ <i>Référence par défaut</i> |
| SCALE | 1 | 1.19444704 | 0.024557 | | | Normal scale parameter |

Lifereg Procedure
Class Level Information

Class Levels Values
AGE 8 1 2 3 4 5 6 7 8

Number of observations used = 1879

Lifereg Procedure

Data Set =WORK.A
Dependent Variable=Log(DUREE)
Censoring Variable=CENS
Censoring Value(s)= 1
Noncensored Values= 1265 Right Censored Values= 614
Left Censored Values= 0 Interval Censored Values= 0

Log Likelihood for LLOGISTIC -2481.700714 ⇨ *Modèle Log Logistique*

Lifereg Procedure

| Variable | DF | Estimate | Std Err | ChiSquare | Pr>Chi | Label/Value |
|----------|----|------------|----------|-----------|--------|-------------------------------------|
| INTERCPT | 1 | 2.69329634 | 0.170331 | 250.0245 | 0.0001 | Intercept |
| AGE | 7 | | | 56.75212 | 0.0001 | |
| | 1 | -0.9097342 | 0.180601 | 25.37386 | 0.0001 | 1 ⇨ <i>Paramètres à transformer</i> |
| | 1 | -0.9306772 | 0.181786 | 26.21072 | 0.0001 | 2 ⇨ <i>transformer</i> |
| | 1 | -0.8451554 | 0.187263 | 20.36906 | 0.0001 | 3 ⇨ |
| | 1 | -0.741975 | 0.189708 | 15.29711 | 0.0001 | 4 ⇨ |
| | 1 | -0.9344953 | 0.191997 | 23.69007 | 0.0001 | 5 ⇨ |
| | 1 | -0.5766439 | 0.213003 | 7.328965 | 0.0068 | 6 ⇨ |
| | 1 | -0.0195009 | 0.240015 | 0.006601 | 0.9352 | 7 ⇨ |
| | 0 | 0 | 0 | . | . | 8 |
| SCALE | 1 | 0.71289653 | 0.016332 | | | Logistic scale parameter |

↳ $\alpha = 1.408$

Lifereg Procedure
Class Level Information

| Class | Levels | Values |
|-------|--------|-----------------|
| AGE | 8 | 1 2 3 4 5 6 7 8 |

Number of observations used = 1879

Lifereg Procedure

Data Set =WORK.A
 Dependent Variable=Log(DUREE)
 Censoring Variable=CENS
 Censoring Value(s)= 1
 Noncensored Values= 1265 Right Censored Values= 614
 Left Censored Values= 0 Interval Censored Values= 0

Log Likelihood for WEIBULL -2545.470743 \Rightarrow *Modèle de Weibull*
 \hookrightarrow n'est pas égal à la log-vraisemblance décrite ici (voir manuel SAS)

Lifereg Procedure

| Variable | DF | Estimate | Std Err | ChiSquare | Pr>Chi | Label/Value |
|----------|----|------------|----------|-----------|--------|--------------------------------|
| INTERCPT | 1 | 3.33821043 | 0.17412 | 367.5614 | 0.0001 | Intercept \Rightarrow -3.296 |
| AGE | 7 | | | 77.01698 | 0.0001 | |
| | 1 | -1.1321535 | 0.183147 | 38.21292 | 0.0001 | 1 \Rightarrow 1.118 |
| | 1 | -1.0666168 | 0.183876 | 33.64871 | 0.0001 | 2 \Rightarrow 1.024 |
| | 1 | -0.872119 | 0.188409 | 21.42638 | 0.0001 | 3 \Rightarrow 0.861 |
| | 1 | -0.8870589 | 0.191704 | 21.41118 | 0.0001 | 4 \Rightarrow 0.876 |
| | 1 | -0.9954031 | 0.192491 | 26.74103 | 0.0001 | 5 \Rightarrow 0.983 |
| | 1 | -0.7022714 | 0.211963 | 10.97717 | 0.0009 | 6 \Rightarrow 0.693 |
| | 1 | -0.0285197 | 0.247474 | 0.013281 | 0.9083 | 7 \Rightarrow 0.028 |
| | 0 | 0 | 0 | . | . | 8 |
| SCALE | 1 | 1.01271776 | 0.022172 | | | Extreme value scale parameter |

$\hookrightarrow \alpha = 0.987$

6. Un estimateur non paramétrique des modèles de durée : Kaplan-Meier

6.1 Présentation générale

L'estimateur de Kaplan Meier est très simple à calculer, et généralise la notion de fonction de répartition empirique en tenant compte des données censurées à droite. C'est pourquoi il sert généralement de base à toute étude sur les durées. Il peut en effet guider le choix d'une forme paramétrique particulière. Rappelons qu'il doit être calculé pour des populations homogènes.

Pour comprendre le principe du calcul, plaçons-nous dans le cas où il n'y a pas de censure. Alors la survie en t peut être simplement estimée par :

$$\hat{S}(t) = 1 - \hat{F}(t) \text{ où } \hat{F}(t) = n_t / N$$

où n_t est nombre de durées inférieures à t et N le nombre total d'observations. Dans SAS, cette fonction de répartition empirique est simplement calculée par une PROC FREQ.

On peut remarquer que la fonction de survie estimée peut s'écrire simplement comme un produit de probabilités conditionnelles. Plaçons nous dans le cas simple sans censure et où on n'observe qu'une seule fois chaque valeur de durée, que l'on notera dans l'ordre croissant t_0, t_1, \dots, t_N , avec $t_0 = 0$. On a alors :

$$S(t) = P(T > t) = \prod_{t_i \leq t} P(T > t_i / T > t_{i-1}) = \prod_{j < i} (1 - q_j)$$

où q_j est la probabilité instantanée de sortir en t_j (l'équivalent de la fonction de hasard en temps discret). Cette probabilité q_j vaut alors $1/(N - j + 1)$, puisqu'on observe une sortie en j parmi les $N - (j - 1)$ personnes qui survivent juste après t_{j-1} . Ces $N - (j - 1)$ personnes sont appelées, par référence aux données médicales, **l'ensemble à risque** en t_j .

Si maintenant certaines durées sont censurées à droite, on va reprendre la même idée, mais en adaptant la notion d'ensemble à risque en t_j . Il sera cette fois défini comme le nombre r_j d'observations ni sorties, ni censurées avant t_j . Alors l'estimateur de q_j s'écrira $1/r_j$, et la survie sera estimée par $\prod_{j < i} (1 - 1/r_j)$.

Dans le cas le plus général où l'on peut observer un nombre d_j supérieur à 1 de sorties à chaque date j , l'estimateur de Kaplan-Meier pour le hasard à la date j sera d_j / r_j , et celui de la survie s'écrira :

$$\hat{S}(t_j) = \prod_{t_j < t} (1 - d_j / r_j)$$

Notons également que l'on peut l'utiliser pour estimer une durée moyenne : puisque l'espérance de la durée peut généralement s'écrire:

$$E(T) = \int_0^{\infty} u f(u) du = \int_0^{\infty} S(u) du$$

on peut utiliser l'estimateur suivant :

$$\hat{T} = \sum_{i=1}^I (t_i - t_{i-1}) \hat{S}(t_i),$$

I étant le nombre de durées différentes observées. La durée moyenne ne sera donc la moyenne empirique que s'il n'y a pas de censure.

Ces estimateurs de la fonction de survie et du hasard sont programmés dans la PROC LIFETEST (voir plus loin pour le détail de sa mise en oeuvre).

L'estimateur de Kaplan Meier a de bonnes propriétés : il est en effet biaisé à distance finie, mais convergent et de loi asymptotique connue (Normale). Il est **donc possible d'utiliser les tests asymptotiques habituels**.

Il est également possible d'utiliser des **méthodes non paramétriques pour tester l'homogénéité de deux sous-populations**. On a vu plus haut que cette homogénéité est essentielle pour interpréter correctement la forme du hasard. SAS fournit, dans la procédure LIFETEST, deux types de tests non paramétriques.

Le premier est un test de rangs généralisant le test de Wilcoxon à des données censurées. Il revient à ordonner l'ensemble des durées T des deux échantillons comparés, en conservant, de plus, l'information sur la censure ($D_i = 1$ si la sortie est observée) et l'échantillon d'origine ($Z_i = 1$ si la durée i vient de l'échantillon 1). On compare alors deux à deux les durées (T_i, T_j) et on attribue un score U_{ij} à toutes ces paires :

$$\begin{cases} U_{ij} = 1 & \text{si } T_i > T_j \text{ et } D_j = 1 \\ U_{ij} = -1 & \text{si } T_i < T_j \text{ et } D_j = 1 \\ U_{ij} = 0 & \text{sinon} \end{cases}$$

On construit alors la statistique de rang $U = \sum_i \sum_{j \neq i} U_{ij} Z_i$. Cela revient à sommer, pour les durées de l'échantillon 1, les scores des paires non censurées. On peut montrer que la loi de U est asymptotiquement normale, de variance connue, sous l'hypothèse nulle du test (homogénéité des deux échantillons, soit même loi de durée (en fait, même loi pour le couple (T_i, D_i)). On rejette l'hypothèse nulle lorsque le rapport $U / \sqrt{V_0(U)}$ dépasse 1,96. On montre également que la statistique de test U s'écrit de façon plus générale :

$$U = \sum_i r(t_i) \left[d_i - \frac{r^1(t_i)}{r(t_i)} \right],$$

où les d_i sont les sorties non censurées en t_i , et $r^1(t_i)$ l'ensemble à risque de l'échantillon 1.

Le second test, dit du « log-rank », revient à comparer les probabilités de sortie des deux échantillons à chaque date t_i . La statistique de test est assez proche de la précédente, puisqu'elle s'écrit :

$$V = \sum_i \left[d_i - \frac{r^1(t_i)}{r(t_i)} \right]$$

Cette statistique est également asymptotiquement normale sous H_0 .

Ces deux types de tests sont effectués dans la PROC LIFETEST. Ils permettent de tester l'homogénéité globale entre strates, mais aussi la significativité d'exogènes particulières. Dans le premier cas, on construit un vecteur Ψ de statistiques de rangs dont les composantes sont définies par :

$$\Psi = \sum_i \sum_{j \neq i} U_{ij} Z_{ik}$$

où Z_{ik} est une variable indicatrice d'appartenance à la strate k .

La statistique globale utilisée pour le premier type d'hypothèse est $\Psi' V^{-1} \Psi$ où V^{-1} est une inverse généralisée de la variance estimée de Ψ qui suit asymptotiquement un $\chi^2(c-1)$ où c est le nombre total de strates. Cette méthode est strictement équivalente aux principes généraux des tests énoncés dans le paragraphe précédent.

6.2 Mise en oeuvre simplifiée.

Cette procédure est utilisable sur des données pouvant être censurées à droite. Elle calcule des fonctions de survie par strates et propose des tests de rang afin d'étudier l'homogénéité des strates.

| | | | | |
|-----------------|---------------|---|---|---------------------------|
| PROC LIFETEST | < Options 1 > | ; | } | Instructions obligatoires |
| TIME variable | < Options 2 > | ; | | |
| By variables | | ; | } | Instructions facultatives |
| ID variables | | ; | | |
| STRATA variable | < Options 3 > | ; | | |
| TEST variables | | ; | | |

Options 1 :

| | |
|--------------------------|---|
| DATA = | précise la table SAS contenant les données. |
| INTERVALS = <i>value</i> | fournit une liste des extrémités des intervalles utilisés dans les calculs de survie. Par défaut, SAS découpe la durée maximale de l'échantillon en dix intervalles. Ainsi, <i>intervals = 5, 10 to 30 by 10</i> produit le découpage [0,5],[5,10],[10,20],[20,30],[30,∞). Elargir l'intervalle |
| METHOD = <i>type</i> | par défaut, SAS utilise les estimateurs de Kaplan Meier de la survie on préférera METHOD = ACT si on veut connaître la fonction de hasard empirique (option conseillée par la suite). |
| NOTABLE | supprime l'impression de la fonction de survie (instruction conseillée sur les fichiers de données individuelles). |

PLOTS = (*type*<,...,&i>type>) produit à la demande les impressions :

| | |
|-----|------------------|
| S | survie empirique |
| LS | -Log(S) |
| LLS | Log(-Log(S)) |
| H | hasard |
| P | densité |

OUTSURV = *data1* crée un fichier SAS contenant différents estimateurs pour chacun des intervalles des différents strates définies par les variables BY et STRATA.

- MIDPOINT, milieu de l'intervalle
- SURVIVAL, survie
- PDF, densité
- HAZARD, hasard.

OUTEST = *data2* crée un fichier contenant les statistiques de rang pour tester les liens entre durées de vie et covariables.

Options 2 :

Variable indique le nom de la variable contenant la durée de vie ; si celle-ci est censurée, elle doit être suivie d'une étoile et du nom de la variable indiquant la censure à droite ; par exemple :

time *t***flag*(1,2) ;

identifie la variable *t*, censurée si la variable *flag* prend les valeurs 1 ou 2.

Options 3 :

La variable STRATA détermine les sous populations sur lesquelles les estimateurs sont calculés. Par rapport à BY, cette instruction permet de réaliser des tests d'homogénéité entre les sous populations. Elle peut être numérique ou alphanumérique. Les données peuvent être formatées dans l'instruction :

STRATA age ;

STRATA age (5 10 20 30) ;

STRATA age (5 to 10) ;

Test:

L'instruction TEST fournit une liste de covariables numériques dont on veut tester les liens avec la durée de vie.

6.3 Exemple d'utilisation

On reprend le fichier précédent en regroupant les classes d'âge (*trage*=1 si *age*=1 ou *age*=2, *trage*=2 sinon).L'appel de la procédure s'écrit:

```
proc lifetest data=a(where=(s='2')) intervals=0 to 24 by 3
method=act plots=(s,ls,h) ;
time duree*cens(1);
strata trage;
```

L'intérêt de l'instruction *interval* est de pouvoir travailler sur des populations plus nombreuses (sorties trimestrielles au lieu de mensuelles) et donc d'obtenir des estimateurs plus précis, dont le profil temporel est aussi souvent plus régulier. Les graphiques sont alors plus lisibles.

The LIFETEST Procedure

Life Table Survival Estimates
TRAGE = 1

| Interval [Lower, Upper) | Number Failed ↳ sorties | Number Censored ↳ censures | Effective Sample Size ↳ taille utile | Conditional Probability of Failure ↳ q _j | Conditional Probability Standard Error | Survival ↳ S | Failure ↳ F | Survival Standard Error |
|----------------------------|-------------------------------|----------------------------------|---|--|---|-----------------|----------------|-------------------------------|
| 0 | 3 | 193 | 824.5 | 0.2341 | 0.0147 | 1.0000 | 0 | 0 |
| 3 | 6 | 181 | 589.0 | 0.3073 | 0.0190 | 0.7659 | 0.2341 | 0.0147 |
| 6 | 9 | 104 | 352.5 | 0.2950 | 0.0243 | 0.5306 | 0.4694 | 0.0178 |
| 9 | 12 | 37 | 201.5 | 0.1836 | 0.0273 | 0.3740 | 0.6260 | 0.0180 |
| 12 | 15 | 40 | 144.5 | 0.2768 | 0.0372 | 0.3053 | 0.6947 | 0.0179 |
| 15 | 18 | 29 | 97.5 | 0.2974 | 0.0463 | 0.2208 | 0.7792 | 0.0172 |
| 18 | 21 | 11 | 63.0 | 0.1746 | 0.0478 | 0.1551 | 0.8449 | 0.0158 |
| 21 | 24 | 9 | 45.0 | 0.2000 | 0.0596 | 0.1281 | 0.8719 | 0.0150 |
| 24 | . | 13 | 22.5 | 0.5778 | 0.1041 | 0.1024 | 0.8976 | 0.0142 |

Evaluated at the Midpoint of the Interval

| Interval [Lower, Upper) | Median Residual Lifetime | Median Standard Error | PDF Standard Error | PDF ↳ densité moyenne | Hazard Standard Error | Hazard ↳ hasard moyen |
|----------------------------|--------------------------------|-----------------------------|--------------------------|--------------------------|-----------------------------|--------------------------|
| 0 | 3 | 6.5855 | 0.3337 | 0.0780 | 0.00492 | 0.08837 |
| 3 | 6 | 5.8287 | 0.3024 | 0.0785 | 0.00508 | 0.12103 |
| 6 | 9 | 7.4221 | 0.5015 | 0.0522 | 0.00464 | 0.115363 |
| 9 | 12 | 7.5442 | 0.6018 | 0.0229 | 0.00357 | 0.067395 |
| 12 | 15 | 6.2733 | 1.4066 | 0.0282 | 0.00413 | 0.107095 |
| 15 | 18 | 8.0666 | 1.3098 | 0.0219 | 0.00381 | 0.116466 |
| 18 | 21 | . | . | 0.00903 | 0.00264 | 0.063768 |
| 21 | 24 | . | . | 0.00854 | 0.00274 | 0.074074 |
| 24 | . | . | . | . | . | . |

Life Table Survival Estimates
TRAGE = 2

| Interval [Lower, Upper) | Number Failed | Number Censored | Effective Sample Size | Conditional Probability of Failure | Conditional Probability Standard Error | Survival | Failure | Survival Standard Error |
|----------------------------|------------------|--------------------|-----------------------------|--|---|----------|---------|-------------------------------|
| 0 | 3 | 221 | 1008.5 | 0.2191 | 0.0130 | 1.0000 | 0 | 0 |
| 3 | 6 | 187 | 721.0 | 0.2594 | 0.0163 | 0.7809 | 0.2191 | 0.0130 |
| 6 | 9 | 100 | 466.5 | 0.2144 | 0.0190 | 0.5783 | 0.4217 | 0.0160 |
| 9 | 12 | 44 | 321.5 | 0.1369 | 0.0192 | 0.4544 | 0.5456 | 0.0167 |
| 12 | 15 | 40 | 251.5 | 0.1590 | 0.0231 | 0.3922 | 0.6078 | 0.0168 |
| 15 | 18 | 21 | 187.5 | 0.1120 | 0.0230 | 0.3298 | 0.6702 | 0.0168 |
| 18 | 21 | 11 | 136.5 | 0.0806 | 0.0233 | 0.2929 | 0.7071 | 0.0167 |
| 21 | 24 | 10 | 103.0 | 0.0971 | 0.0292 | 0.2693 | 0.7307 | 0.0168 |
| 24 | . | 14 | 50.0 | 0.2800 | 0.0635 | 0.2431 | 0.7569 | 0.0171 |

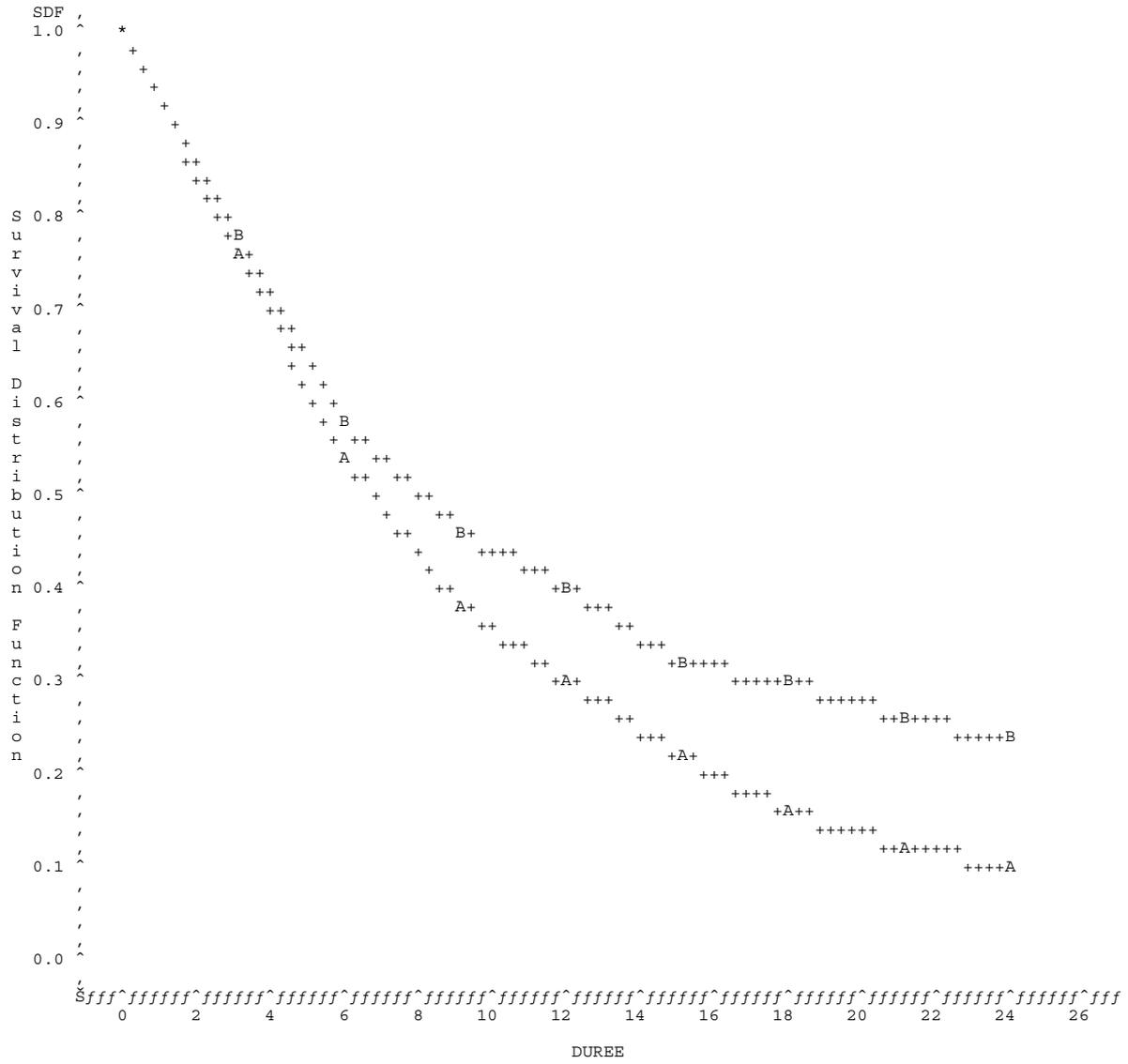
Evaluated at the Midpoint of the Interval

| Interval [Lower, Upper) | Median Residual Lifetime | Median Standard Error | PDF | PDF Standard Error | Hazard | Hazard Standard Error | |
|----------------------------|--------------------------------|-----------------------------|--------|--------------------------|---------|-----------------------------|----------|
| 0 | 3 | 7.8956 | 0.3810 | 0.0730 | 0.00434 | 0.082034 | 0.005476 |
| 3 | 6 | 9.0841 | 0.6993 | 0.0675 | 0.00440 | 0.099336 | 0.007183 |
| 6 | 9 | 12.4701 | 1.7018 | 0.0413 | 0.00384 | 0.080032 | 0.007945 |
| 9 | 12 | . | . | 0.0207 | 0.00300 | 0.048971 | 0.007363 |
| 12 | 15 | . | . | 0.0208 | 0.00314 | 0.057595 | 0.009073 |
| 15 | 18 | . | . | 0.0123 | 0.00261 | 0.039548 | 0.008615 |
| 18 | 21 | . | . | 0.00787 | 0.00232 | 0.02799 | 0.008432 |
| 21 | 24 | . | . | 0.00871 | 0.00267 | 0.034014 | 0.010742 |
| 24 | . | . | . | . | . | . | . |

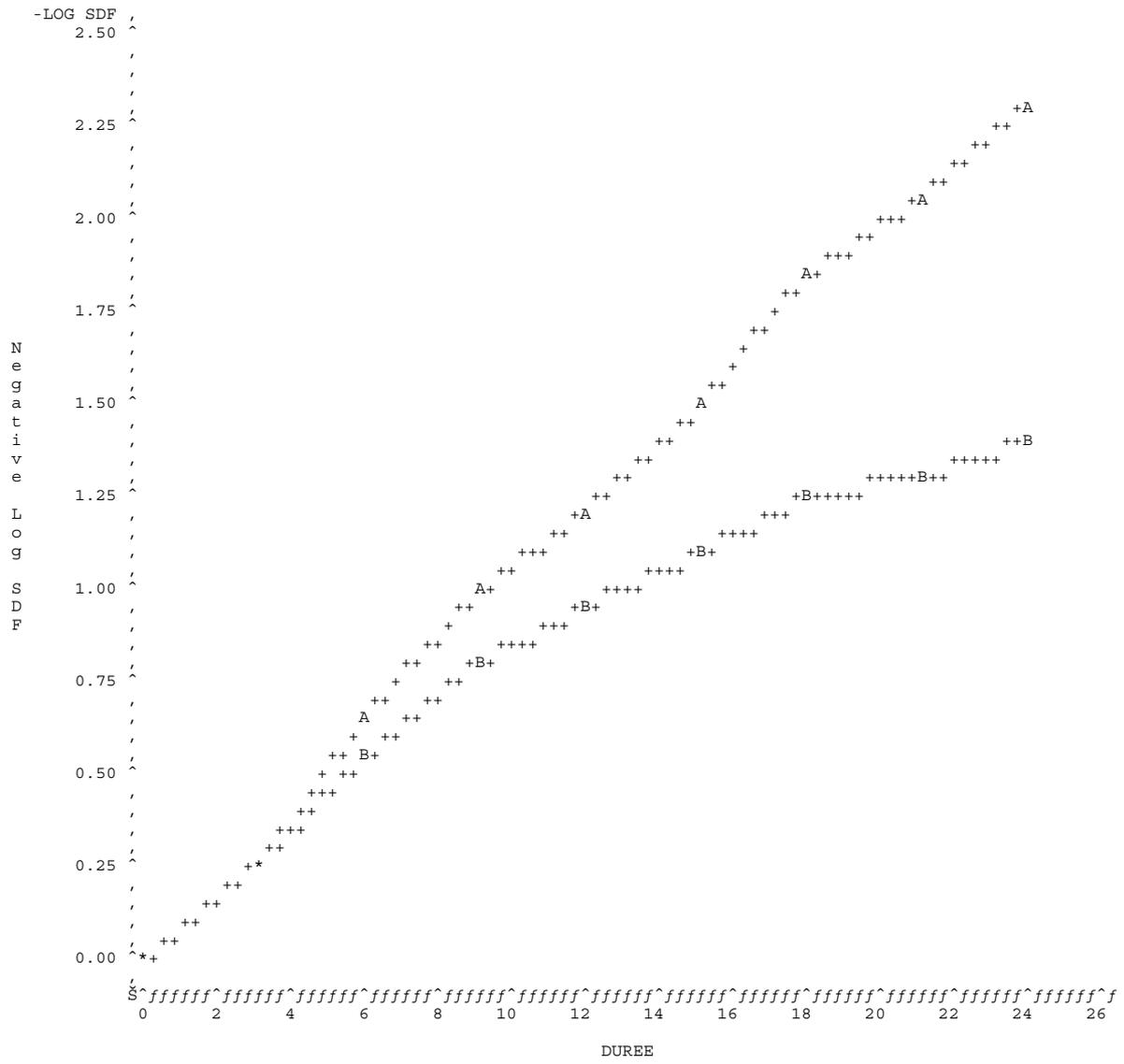
Summary of the Number of Censored and Uncensored Values

| TRAGE | Total | Failed | Censored | %Censored |
|-------|-------|--------|----------|-----------|
| 1 | 843 | 617 | 226 | 26.8090 |
| 2 | 1036 | 648 | 388 | 37.4517 |
| Total | 1879 | 1265 | 614 | 32.6770 |

Survival Function Estimates



-Log(Survival Function) Estimates



The LIFETEST Procedure

Testing Homogeneity of Survival Curves over Strata
Time Variable DUREE

Rank Statistics

↳ Statistiques de rang par sous-groupe

| TRAGE | Log-Rank | Wilcoxon |
|-------|----------|----------|
| 1 | 87.366 | 63153 |
| 2 | -87.366 | -63153 |

↳ opposées car il n'y a que 2 sous groupes

Covariance Matrix for the Log-Rank Statistics

↳ Variance estimée du premier vecteur de statistiques de rang (log rank)

| TRAGE | 1 | 2 |
|-------|----------|----------|
| 1 | 275.349 | -275.349 |
| 2 | -275.349 | 275.349 |

Covariance Matrix for the Wilcoxon Statistics

↳ Variance estimée du vecteur de statistiques de Wilcoxon

| TRAGE | 1 | 2 |
|-------|----------|----------|
| 1 | 4.3581E8 | -4.358E8 |
| 2 | -4.358E8 | 4.3581E8 |

Test of Equality over Strata

↳ valeur test pour les deux tests de rang

| Test | Chi-Square | DF | Pr > |
|-----------|------------|----|--------|
| Log-Rank | 27.7207 | 1 | 0.0001 |
| Wilcoxon | 9.1514 | 1 | 0.0025 |
| -2Log(LR) | 37.8895 | 1 | 0.0001 |

● Valeur des tests pour les deux tests de rang ↔ Proba faibles: rejet de l'hypothèse

● Test rapport vraisemblance ↔ d'homogénéité

↳ H0: les lois sont exponentielles de même paramètre dans les deux sous-groupes

↳ H1: les lois sont exponentielles de paramètre différent

(hypothèses plus restrictives que les deux tests précédents)

7. Une estimation semi-paramétrique : le modèle de Cox

7.1 Présentation générale

Une méthode d'estimation semi-paramétrique est disponible dans la PROC PHREG de SAS. Elle concerne les modèles à hasard proportionnels présentés dans la partie 5.3 avec la spécification suivante pour la fonction de hasard :

$$h(t) = \exp(Xb)h_0(t),$$

où h_0 . Elle repose sur la maximisation de la « vraisemblance partielle » de Cox. Elle présente en outre l'avantage de ne pas contraindre les variables explicatives à être constantes au cours du temps.

7.2 Vraisemblance partielle de Cox

Reprenons la situation la plus simple où l'on observe autant de durées que d'individus et où il n'y a pas de censure; on ordonne les valeurs des I durées différentes observées: $t_1 < t_2 < \dots < t_I$. Soit comme précédemment $r(t_i)$ l'ensemble à risque en t_i . La probabilité pour que ce soit l'individu j de $r(t_i)$ qui sorte en t_i vaut :

$$\frac{h_0(t_i) \exp(X_j b)}{\sum_{k \in r(t_i)} h_0(t_i) \exp(X_k b)}$$

Le dénominateur est la probabilité qu'une sortie ait lieu en t_i au sein de l'ensemble à risque. Il vaut la somme des probabilités de sortie de tous les individus de cet ensemble. L'expression se simplifie puisque $h_0(t_i)$ figure dans le dénominateur et le numérateur, et elle vaut finalement:

$$\frac{\exp(X_j b)}{\sum_{k \in r(t_i)} \exp(X_k b)}$$

sachant que c'est l'individu j qui sort à la date i . La vraisemblance partielle de Cox est le produit de ces probabilités pour l'ensemble des sorties.

$$L(b) = \prod_{i=1}^I \frac{\exp(X_{j_i} b)}{\sum_{k \in r(t_i)} \exp(X_k b)}$$

S'il n'y a pas de censure, elle s'interprète comme la vraisemblance de la statistique de rang associée aux durées. L'estimateur semi-paramétrique de b va être obtenu en maximisant la log-vraisemblance partielle par rapport à b au moyen d'une méthode itérative. L'estimateur obtenu converge presque sûrement vers b et est asymptotiquement normal.

7.3 Estimation non paramétrique du hasard de base

Plutôt que le hasard de base, on préfère, en général, estimer directement sa fonction de survie. Dans le modèle de Cox, on a vu que la fonction de survie s'écrivait :

$$S(t) = [S_0(t)]^{\exp(Xb)}$$

Cette relation découle de la définition du modèle et de la relation générale entre hasard et survie. Kableish et Prentice en déduisent une méthode d'estimation de la « survie de base » en deux étapes. Dans une première étape, on estime b par une maximisation de vraisemblance partielle, comme décrit précédemment en 7.2. Ensuite, b étant remplacé par son estimation issue de la première étape, on maximise la vraisemblance par rapport à S_0 .

Cette procédure revient à estimer la survie de base par :

$$\hat{S}_0(t) = \prod_{t_i < t} \hat{\alpha}_i$$

où

$$\hat{\alpha}_i \exp(X_i \hat{b}) = 1 - \frac{\exp(X_i \hat{b})}{\sum_{k \in r(t_i)} \exp(X_k \hat{b})}$$

L'estimateur utilisé dans PHREG est celui de Breslow:

$$\hat{S}_0(t) = \prod_{t_i < t} 1 - \frac{d_i}{\sum_{k \in r(t_i)} \exp(X_k \hat{b})}$$

où, comme précédemment, d_i est une variable muette valant 1 si la durée n'est pas censurée. Le « hasard intégré » $\hat{H}_0(t) = \int_0^t h_0(u) du$ est alors simplement estimé par $-\log(\hat{S}_0(t))$.

L'estimateur \hat{b} a des propriétés de convergence presque sûre et de normalité asymptotique. Cela permet d'effectuer des tests asymptotiques sur les paramètres, comme dans les modèles pleinement paramétriques.

L'estimation de la vraisemblance de Cox repose de manière cruciale sur l'hypothèse de hasard proportionnel. Cette hypothèse peut être confirmée qualitativement par des contrôles graphiques. En effet, considérons un modèle simple avec pour seule variable exogène une constante, et donc s'écrivant :

$$h(t) = \exp(b)h_0(t)$$

La relation sur le hasard intégré s'écrira alors :

$$H(t) = \exp(b)H_0(t)$$

d'où $\log(H(t)) - \log(H_0(t)) = b$. L'écart entre les deux courbes de hasard intégré est donc constant. De manière générale, on trouvera un écart constant entre les divers groupes définis par les valeurs des exogènes si l'hypothèse de hasard proportionnel est vérifiée. Il existe également des tests paramétriques pour la spécification proportionnelle (Voir « Pour en savoir plus », en particulier MOREAU).

7.4 Mise en oeuvre simplifiée.

La procédure PHREG est utilisable sur des données non censurées ou censurées à droite. Elle calcule un estimateur non paramétrique du hasard de base et des estimateurs paramétriques des coefficients associés aux covariables affectant le hasard de base sous la forme $\exp(Xb)$.

```
PROC PHREG < Options 1 > ;
MODEL      duree * flag = exogènes;
```



```
FREQ      variable;
```

```
OUTPUT    < Options 2 > ;
```

```
BASELINE  < Options 3 > ;
```

Options 1 :

```
DATA=
OUTEST=data1      nom du data qui contiendra les estimateurs des coefficients des covariables
COVOUT            ajoute dans OUTEST la matrice de variance-covariance
```

Options 2 :

```
OUT= data2      nom du data de sortie construit à partir du tableau initial et contenant les
                 statistiques requises, parmi lesquelles les plus utiles sont :
```

```
XBETA =         Xb
SURVIVAL        survie
LOGSURV         Log(survie)
```

Options 3 :

```
OUT= data3      nom du data de sortie contenant la valeur de la survie pour les valeurs
                 possibles de duree
```

```
COVARIATES= data nom d'un data contenant des valeurs particulières des covariables pour
                 lesquelles on cherche à calculer la survie ; cette option est purement
                 illustrative. En son absence, SAS constitue un individu fictif pour lequel les
                 valeurs de X sont les valeurs moyennes du fichier des durées.
```

```
XBETA =         Xb
SURVIVAL        survie
LOGSURV         Log(survie)
```

7.5 Exemples d'utilisation

On reprend le même fichier en créant une variable jeune muette valant 1 si *trage*=1. L'appel de la procédure s'écrit:

```
data cov;
** on veut obtenir dans b le hasard de base pour les deux sous
   populations jeunes=0 et jeunes=1 **;
input jeunes;cards;
1
0
;

proc phreg data=a(where=(s='2'));
model duree*cens(1)=jeunes;
baseline out=b covariates=cov survival=s logsurv=ls;
```

```
proc print data=b;
```

The PHREG Procedure

Data Set: WORK.A
 Dependent Variable: DUREE
 Censoring Variable: CENS
 Censoring Value(s): 1
 Ties Handling: BRESLOW

Summary of the Number of
 Event and Censored Values

| Total | Event | Censored | Percent Censored |
|-------|-------|----------|------------------|
| 1879 | 1265 | 614 | 32.68 |

Testing Global Null Hypothesis: BETA=0

| Criterion | Without Covariates | With Covariates | Model Chi-Square |
|----------------|--------------------|-----------------|-----------------------------|
| -2 LOG L Score | 17280.160 | 17255.429 | 24.731 with 1 DF (p=0.0001) |
| Wald | . | . | 25.022 with 1 DF (p=0.0001) |
| | | | 24.862 with 1 DF (p=0.0001) |

Analysis of Maximum Likelihood Estimates

| Variable | DF | Parameter Estimate | Standard Error | Wald Chi-Square | Pr > Chi-Square | Risk Ratio |
|----------|----|--------------------|----------------|-----------------|-----------------|------------|
| JEUNES | 1 | 0.281880 | 0.05653 | 24.86154 | 0.0001 | 1.326 |

↳ \hat{b} translate de $\exp(0.282)$
 le hasard de base entre les jeunes et les autres

| OBS | JEUNES | DUREE | S | LS \Rightarrow data b |
|---|--------|-------|---------|---|
| <i>Calcul de la survie et de son logarithme pour trois sous populations (jeunes=0 ou 1 et défaut)</i> | | | | |
| 1 | 1 | 0 | 1.00000 | 0.00000 \Rightarrow Jeunes=1 |
| 2 | 1 | 1 | 0.85858 | -0.15247 |
| 3 | 1 | 2 | 0.74676 | -0.29201 |
| 4 | 1 | 3 | 0.65740 | -0.41947 |
| 5 | 1 | 4 | 0.57374 | -0.55557 |
| 6 | 1 | 5 | 0.51060 | -0.67218 |
| 7 | 1 | 6 | 0.45832 | -0.78019 |
| 8 | 1 | 7 | 0.40159 | -0.91233 |
| 9 | 1 | 8 | 0.36805 | -0.99953 |
| 10 | 1 | 9 | 0.34478 | -1.06486 |
| 11 | 1 | 10 | 0.32130 | -1.13537 |
| 12 | 1 | 11 | 0.30285 | -1.19453 |
| 13 | 1 | 12 | 0.26811 | -1.31638 |
| 14 | 1 | 13 | 0.24364 | -1.41204 |
| 15 | 1 | 14 | 0.23292 | -1.45707 |
| 16 | 1 | 15 | 0.21636 | -1.53080 |
| 17 | 1 | 16 | 0.19889 | -1.61499 |
| 18 | 1 | 17 | 0.18507 | -1.68705 |
| 19 | 1 | 18 | 0.17081 | -1.76718 |
| 20 | 1 | 19 | 0.16657 | -1.79231 |
| 21 | 1 | 20 | 0.16195 | -1.82045 |
| 22 | 1 | 21 | 0.15705 | -1.85117 |
| 23 | 1 | 22 | 0.14802 | -1.91040 |
| 24 | 1 | 23 | 0.13730 | -1.98560 |
| 25 | 1 | 24 | 0.12603 | -2.07121 |
| 26 | 1 | 25 | 0.12025 | -2.11816 |
| 27 | 1 | 26 | 0.11352 | -2.17581 |
| 28 | 1 | 27 | 0.10898 | -2.21657 |
| 29 | 1 | 28 | 0.10653 | -2.23934 |
| 30 | 1 | 29 | 0.10126 | -2.29011 |
| 31 | 1 | 30 | 0.09553 | -2.34828 |
| 32 | 1 | 31 | 0.09249 | -2.38068 |
| 33 | 1 | 34 | 0.08654 | -2.44710 |
| 34 | 1 | 35 | 0.07979 | -2.52840 |
| 35 | 1 | 36 | 0.07204 | -2.63055 |
| 36 | 0 | 0 | 1.00000 | 0.00000 \Rightarrow Autres (jeunes=0) |
| 37 | 0 | 1 | 0.89135 | -0.11502 |
| 38 | 0 | 2 | 0.80229 | -0.22028 |
| 39 | 0 | 3 | 0.72875 | -0.31643 |
| 40 | 0 | 4 | 0.65764 | -0.41910 |
| 41 | 0 | 5 | 0.60226 | -0.50707 |
| 42 | 0 | 6 | 0.55513 | -0.58855 |
| 43 | 0 | 7 | 0.50246 | -0.68823 |
| 44 | 0 | 8 | 0.47048 | -0.75401 |
| 45 | 0 | 9 | 0.44785 | -0.80329 |
| 46 | 0 | 10 | 0.42465 | -0.85648 |
| 47 | 0 | 11 | 0.40612 | -0.90111 |
| 48 | 0 | 12 | 0.37045 | -0.99303 |
| 49 | 0 | 13 | 0.34466 | -1.06520 |
| 50 | 0 | 14 | 0.33315 | -1.09916 |
| 51 | 0 | 15 | 0.31513 | -1.15478 |
| 52 | 0 | 16 | 0.29574 | -1.21829 |
| 53 | 0 | 17 | 0.28009 | -1.27265 |
| 54 | 0 | 18 | 0.26366 | -1.33310 |
| 55 | 0 | 19 | 0.25871 | -1.35205 |
| 56 | 0 | 20 | 0.25327 | -1.37328 |
| 57 | 0 | 21 | 0.24747 | -1.39645 |
| 58 | 0 | 22 | 0.23666 | -1.44114 |
| 59 | 0 | 23 | 0.22361 | -1.49787 |
| 60 | 0 | 24 | 0.20962 | -1.56245 |

| OBS | JEUNES | DUREE | S | LS |
|-----|---------|-------|---------|----------------------------|
| 61 | 0.00000 | 25 | 0.20233 | -1.59787 |
| 62 | 0.00000 | 26 | 0.19372 | -1.64135 |
| 63 | 0.00000 | 27 | 0.18785 | -1.67210 |
| 64 | 0.00000 | 28 | 0.18465 | -1.68928 |
| 65 | 0.00000 | 29 | 0.17771 | -1.72758 |
| 66 | 0.00000 | 30 | 0.17008 | -1.77146 |
| 67 | 0.00000 | 31 | 0.16598 | -1.79590 |
| 68 | 0.00000 | 34 | 0.15787 | -1.84601 |
| 69 | 0.00000 | 35 | 0.14848 | -1.90734 |
| 70 | 0.00000 | 36 | 0.13746 | -1.98440 |
| 71 | 0.44864 | 0 | 1.00000 | 0.00000 ⇨ $X=0.45$ |
| 72 | 0.44864 | 1 | 0.87764 | -0.13052 ⇨ Moyenne de X |
| 73 | 0.44864 | 2 | 0.77882 | -0.24998 ⇨ dans |
| 74 | 0.44864 | 3 | 0.69831 | -0.35909 ⇨ l'échantillon |
| 75 | 0.44864 | 4 | 0.62151 | -0.47560 ⇨ (45% de jeunes) |
| 76 | 0.44864 | 5 | 0.56247 | -0.57542 ⇨ par défaut |
| 77 | 0.44864 | 6 | 0.51279 | -0.66789 |
| 78 | 0.44864 | 7 | 0.45794 | -0.78101 |
| 79 | 0.44864 | 8 | 0.42501 | -0.85565 |
| 80 | 0.44864 | 9 | 0.40189 | -0.91158 |
| 81 | 0.44864 | 10 | 0.37835 | -0.97194 |
| 82 | 0.44864 | 11 | 0.35966 | -1.02259 |
| 83 | 0.44864 | 12 | 0.32404 | -1.12689 |
| 84 | 0.44864 | 13 | 0.29856 | -1.20879 |
| 85 | 0.44864 | 14 | 0.28727 | -1.24734 |
| 86 | 0.44864 | 15 | 0.26970 | -1.31045 |
| 87 | 0.44864 | 16 | 0.25094 | -1.38252 |
| 88 | 0.44864 | 17 | 0.23593 | -1.44421 |
| 89 | 0.44864 | 18 | 0.22029 | -1.51281 |
| 90 | 0.44864 | 19 | 0.21560 | -1.53432 |
| 91 | 0.44864 | 20 | 0.21047 | -1.55841 |
| 92 | 0.44864 | 21 | 0.20501 | -1.58471 |
| 93 | 0.44864 | 22 | 0.19487 | -1.63542 |
| 94 | 0.44864 | 23 | 0.18272 | -1.69979 |
| 95 | 0.44864 | 24 | 0.16981 | -1.77308 |
| 96 | 0.44864 | 25 | 0.16312 | -1.81327 |
| 97 | 0.44864 | 26 | 0.15526 | -1.86262 |
| 98 | 0.44864 | 27 | 0.14994 | -1.89752 |
| 99 | 0.44864 | 28 | 0.14705 | -1.91701 |
| 100 | 0.44864 | 29 | 0.14079 | -1.96047 |
| 101 | 0.44864 | 30 | 0.13395 | -2.01027 |
| 102 | 0.44864 | 31 | 0.13029 | -2.03800 |
| 103 | 0.44864 | 34 | 0.12309 | -2.09486 |
| 104 | 0.44864 | 35 | 0.11481 | -2.16446 |
| 105 | 0.44864 | 36 | 0.10520 | -2.25191 |

8. La sélection endogène

8.1 position du problème

Jusqu'ici, on a considéré que les modèles de durée étaient estimés sur des fichiers d'épisodes, décrivant par exemple des durées de chômages. C'est la façon la plus naturelle d'aborder ce type de problème. Elle consiste à réaliser un échantillon d'individus entrant au chômage au cours d'un intervalle de temps, et à les suivre jusqu'à leur sortie, ou à défaut jusqu'à la fin de la période de collecte. On dispose alors d'un fichier retraçant les caractéristiques d'un ensemble d'épisodes complets ou censurés de chômage de durées différentes. On parle alors de fichier de flux (flow sampling).

Dans la pratique, que ce soit pour des raisons de coût ou de commodité, une technique de collecte plus simple est fréquemment utilisée. Elle consiste à tirer à une date donnée un échantillon de chômeurs dans la population totale, et à les suivre éventuellement pendant une certaine durée. Certains vont sortir du chômage, d'autres non. On appelle un tel tirage un fichier de stock (stock sampling) puisqu'il est constitué du stock de chômeurs présents à une date donnée. Si le mode de constitution est plus aisé que pour un fichier de flux, le traitement statistique est beaucoup plus complexe, si l'on veut éviter des biais importants dans l'estimation des durées. On peut ainsi montrer que pour un processus de renouvellement et une loi des durées exponentielle, ne pas corriger les estimateurs conduit à un **doublément** des durées estimées (Lancaster (1991)). Un exemple simple permet de s'en convaincre, et de fournir des méthodes d'estimation convergentes dans le cas des procédures non paramétriques.

Pour comprendre les difficultés liées à l'estimation des taux de sortie dans les fichiers de stock, il faut distinguer les différentes générations de chômeurs en fonction de leurs dates d'arrivée au chômage. Supposons pour simplifier que les épisodes de chômage durent au plus trois périodes (la généralisation à un nombre plus élevé est immédiate). Il y a alors trois générations : ceux qui viennent d'arriver au chômage, ceux qui sont arrivés à la date précédente, et ceux qui sont arrivés il y a deux périodes. Le fichier constitué par le stock de chômeurs au moment de l'entrée en chômage de la génération 1 contient alors les populations suivantes :

| Génération | Durée | | |
|------------|------------|------------|------------|
| | $\tau = 1$ | $\tau = 2$ | $\tau = 3$ |
| $K = 1$ | n_{11} | n_{12} | n_{13} |
| $K = 2$ | 0 | n_{22} | n_{23} |
| $K = 3$ | 0 | 0 | n_{33} |

où $n_{k\tau}$ est le nombre de sortants du chômage après la durée τ pour la génération k . Dans le tableau, la stratification s'effectue donc selon deux critères, l'un exogène, la génération, l'autre endogène, la durée. Les chômeurs de générations anciennes, mais dont les durées sont brèves sont sorties de l'état, donc absentes dans l'échantillon, d'où la présence de zéros dans le tableau. La population de référence des chômeurs, si l'on disposait d'un fichier de flux, se répartirait de la façon suivante dans l'échantillon :

| Génération | Durée | | |
|------------|------------|------------|------------|
| | $\tau = 1$ | $\tau = 2$ | $\tau = 3$ |
| $K = 1$ | n_{11} | n_{12} | n_{13} |
| $K = 2$ | n_{21} | n_{22} | n_{23} |
| $K = 3$ | n_{31} | n_{32} | n_{33} |

Par rapport au fichier de flux, le fichier de stock résulte d'un tirage selon une stratification partiellement endogène ; il est donc non représentatif de la population sous-jacente. L'utilisation de procédures standard d'estimation entraîne donc des biais, liés à cette sélection endogène.

Supposons, comme dans la plupart des cas, que la loi de la durée ne dépende pas de la date d'entrée au chômage, et donc de la génération. Sur l'échantillon complet, la probabilité de sortir du chômage à la date j s'écrit :

$$p_j = \frac{n_{1j} + n_{2j} + n_{3j}}{\sum_{k,j} n_{kj}}$$

et les taux de sortie instantanés (hasards) :

$$q_j = \frac{p_j}{\sum_{l=j}^3 p_l}$$

Estimée sans précaution sur le fichier de stock, la probabilité de sortir du chômage à la date j serait :

$$p_j^* = \frac{\sum_{k=1}^j n_{kj}}{\sum_j \sum_{k=1}^j n_{kj}}$$

et les taux de sortie instantanés :

$$q_j^* = \frac{p_j^*}{\sum_{l=j}^3 p_l^*}$$

Le biais va provenir de l'omission des sorties non observées pour les générations anciennes. Il conduit à sous estimer les taux de sortie et à faire apparaître un profil selon j plus croissant (ou moins décroissant) qu'il n'est en réalité. La raison tient au fait que pour les durées courtes, seules les générations récentes sont prises en compte. On observe donc moins de chômeurs de courte durée que

dans la population de référence, par conséquent les taux de sortie sont sous estimés, et ceci d'autant plus que les durées de chômage sont brèves.

Deux approches sont envisageables pour redresser le biais lié à la sélection endogène.

↳ **Si l'on dispose d'une information** sur la taille des différentes générations de chômeurs, le biais peut être redressé au moyen d'une surpondération des durées courtes. Si l'on sait par exemple que toutes les générations ont la même taille, il suffit de pondérer n_{11} par 3, n_{12} et n_{22} par 3/2 pour que les estimateurs p_j^* et q_j^* deviennent sans biais.

↳ **Si l'on ne dispose pas d'information** sur la taille des générations de chômeurs, et que l'on ne souhaite pas faire d'hypothèse, on peut adopter une méthode fondée sur le maximum de vraisemblance conditionnelle. Cette procédure part du fait que *conditionnellement à l'appartenance à une génération*, on peut trouver des estimateurs sans biais des taux de sortie instantanés. En particulier, la première génération, observée sur la totalité de la durée, fournit les estimateurs requis. C'est aussi le cas de la seconde à partir de la date 2, etc... En prenant en compte, la totalité du fichier, on peut montrer que les meilleurs estimateurs fondés sur la vraisemblance conditionnelle s'écrivent :

$$\hat{q}_1 = \frac{n_{11}}{n_{11} + n_{12} + n_{13}}$$

$$\hat{q}_2 = \frac{n_{12} + n_{22}}{n_{12} + n_{13} + n_{22} + n_{23}}$$

Ces résultats sont assez intuitifs car ils prennent en compte les plus grands populations possibles pour fonder les estimateurs, sans que celles-ci soient touchées par la sélection endogène ; on en déduit :

$$\hat{p}_1 = \hat{q}_1$$

$$\hat{p}_2 = (1 - \hat{q}_1)\hat{q}_2$$

8.2 les méthodes d'estimation paramétrique sur fichiers de stock

Gouriéroux et Monfort (1991) proposent des méthodes d'estimation nécessitant un ensemble d'hypothèses parfois moins contraignant que le cadre précédent, et prenant en compte l'influence de divers facteurs explicatifs.

Soit, pour un individu donné x , la valeur des variables exogènes, y la valeur de la durée totale et t la valeur de la durée écoulée depuis la date d'entrée au chômage. Faisons l'hypothèse que la loi de la durée Y est indépendante de celle de la durée écoulée T , conditionnellement aux variables explicatives x . Cette hypothèse est d'importance car elle suppose une permanence du phénomène considéré. En particulier concernant le chômage, elle interdit que la loi conditionnelle de la durée dépende de l'état du marché du travail. Ce dernier ne peut être introduit que sous forme de facteur explicatif. De Toldi, Gouriéroux et Monfort (1992) relâchent cette hypothèse assez usuelle, mais sur des fichiers de flux, afin de mettre en évidence des effets saisonniers.

Les observations de (X, Y, T) sont indépendantes d'un individu à l'autre et de même loi ; du fait de l'hypothèse d'indépendance conditionnelle, celle-ci peut se décomposer en une expression de la forme :

$$f_0(x)g_0(t/x)f(y/x;\theta_0)$$

où θ_0 désigne la valeur recherchée des paramètres.

Les observations étant menées conditionnellement au fait que $Y > T$, la loi admet la densité :

$$l_0(x, y, t) = \frac{f_0(x)g_0(t/x)f(y/x;\theta_0)\mathbf{I}_{y>t}}{\iiint_{y>t} f_0(x)g_0(t/x)f(y/x;\theta_0)dydxdt}$$

Si on note $S(y/x;\theta_0)$ la fonction de survie associée à Y , l'expression précédente peut s'écrire :

$$l_0(x, y, t) = \frac{f_0(x)g_0(t/x)f(y/x;\theta_0)\mathbf{I}_{y>t}}{\iint f_0(x)g_0(t/x)S(t/x;\theta_0)dxdt}$$

La sélection endogène entraîne une particularité : les variables qui sont exogènes lorsqu'on s'intéresse à la population ne le sont plus sur fichier de stock (dans le cas général où les variables X et Y ne sont pas indépendantes). En effet, la loi marginale de X sur les données de l'échantillon s'écrit :

$$l_0(x) = \frac{f_0(x) \int g_0(t/x)S(t/x;\theta_0)dt}{\iint f_0(x)g_0(t/x)S(t/x;\theta_0)dxdt}$$

Cette densité dépend de θ_0 au travers de S . Ne pas tenir compte de cette loi marginale entraînera donc au moins une perte d'information sur le paramètre θ_0 .

Si l'on raisonne conditionnellement à x , la densité s'écrit :

$$l(y, t) = \frac{g_0(t)f(y)\mathbf{I}_{y>t}}{\int g_0(t)S(t)dt}$$

en allégeant les notations liées au conditionnement.

Si l'épisode est censuré, cette densité se transforme en probabilité :

$$L(y, t) = \frac{g_0(t)S(y)}{\int g_0(t)S(t)dt}$$

Comme dans la partie précédente, la résolution dépend des informations dont on dispose sur l'effet de génération $g_0(t)$.

↪ **En environnement stationnaire**, toutes les générations ont la même taille et la densité se simplifie en :

$$l(y) = \frac{f(y)}{\int S(t)dt} = \frac{f(y)}{\mu}$$

où μ est l'espérance de la durée Y . On est dans la situation d'un processus de renouvellement. Un cas particulier intéressant est celui où le fichier de stock est une coupe instantanée, dans laquelle seule

les anciennetés sont disponibles et où l'on ne réalise pas de suivi. Toutes les durées sont alors censurées en $y = t$. La vraisemblance s'écrit :

$$L(t) = \frac{S(t)}{\mu}$$

Un cas particulier du cas particulier est celui de la loi exponentielle. En utilisant la formule ci-dessus, la loi des durées censurées est elle aussi exponentielle, et permet d'ajuster un modèle sans difficulté, avec les logiciels standards.

↪ **En situation non stationnaire**, les générations ne sont pas de taille identique. Il est rare de disposer d'information conditionnellement aux variables X . En général, seules des informations globales résultant d'un suivi macro-économique sont disponibles. Certains auteurs font cependant des hypothèses de séparabilité du type (Nickell) :

$$g_0(t) = g(t)c(x)$$

afin de procéder à des repondérations visant à fournir des estimateurs sans biais. Notons que si la fonction de répartition de l'effet de génération $G_0(t)$ était connue de façon extérieure, la solution la plus simple consisterait à adopter une autre approche visant à repondérer les observations ; on utiliserait alors l'estimateur redressé défini comme la solution de :

$$\text{Max}_{\theta} \sum_{i=1}^n \frac{1}{G_0(y_i / x_i)} \log f(y_i / x_i; \theta)$$

qui est convergent et sans biais.

Dans le cas où aucune information extérieure sur l'effet de génération n'est disponible, on peut adopter une méthode de maximum de vraisemblance conditionnel. On fonde l'estimateur sur la loi de Y sachant T , loi qui doit être calculée sur l'échantillon :

$$l_0(y / t) = \frac{f(y)I_{y>t}}{S(t)}$$

Cette loi conditionnelle ne dépend des vraies lois inconnues f_0, g_0, θ_0 que par l'intermédiaire de θ_0 . Elle peut donc être utilisée comme base d'une procédure d'estimation. L'estimateur du maximum de vraisemblance conditionnel est défini comme la solution du problème :

$$\text{Max}_{\theta} \sum_{i=1}^n \log \frac{f(y_i / x_i; \theta)}{S(t_i / x_i; \theta)}$$

Cette démarche nécessite une information plus riche dans le fichier que celle du maximum de vraisemblance global avec repondération. En particulier, un suivi temporel de individus est indispensable. En effet, si toutes les durées étaient censurées en t (cas de la coupe instantanée), la forme précédente serait inopérante.

8.4 un exemple d'estimation dans le cas du modèle de Weibull

Dans le cas d'un modèle de durée sans sélection endogène, on déduit de la formule de la partie 5.3 l'expression de la log-vraisemblance :

$$\log(L) = \sum_{i=1}^n d_i \log(h(y_i)) + \sum_{i=1}^n \log(S(y_i))$$

où y_i est la variable de durée et d_i vaut 1 si la durée n'est pas censurée.

Si l'on s'intéresse à une modèle de Weibull, la log-vraisemblance devient :

$$\log(L) = \sum_{i=1}^n d_i [\log(\alpha) + X_i b + (\alpha - 1) \log(y_i)] + \sum_{i=1}^n y_i^\alpha \exp(X_i b)$$

En présence de sélection endogène, on a vu qu'une méthode convergente consiste à rendre maximale la vraisemblance conditionnelle ; celle-ci s'écrit simplement :

$$\log(L) = \sum_{i=1}^n d_i [\log(\alpha) + X_i b + (\alpha - 1) \log(y_i)] + \sum_{i=1}^n (y_i^\alpha - t_i^\alpha) \exp(X_i b)$$

où t_i est la durée déjà écoulée au moment de la constitution du fichier de stock. t_i vaut 0 si l'épisode n'est pas soumis à sélection endogène ; c'est par exemple le cas lorsqu'il débute après la constitution du fichier. Formellement, la vraisemblance conditionnelle n'est pas très différente de la vraisemblance totale. Néanmoins, les procédures standard ne permettent pas d'estimer de tels modèles. Il faut écrire soi-même la maximisation de la log-vraisemblance.

Cependant, dans SAS, on dispose de la procédure NLIN dont une propriété (parmi d'autres) est de permettre la minimisation de fonction, si l'on introduit sa forme fonctionnelle et celle de ses dérivées. Cette procédure doit être paramétrée d'une façon particulière (se reporter à la brochure SAS) :

- le paramètre SIGSQ doit être rendu égal à 1
- l'instruction MODEL doit être désactivée pour générer un résidu de 1
- la fonction `_loss_` doit être rendue égale à l'opposé de la log-vraisemblance
- les paramètres `der.xxx` doivent pour leur part correspondre aux dérivées de la log-vraisemblance (et non à leur opposé).

On a fabriqué un exemple en simulant de façon approchée un fichier de stock à partir du calendrier de l'enquête sur l'emploi (à cet effet on ne conserve que les épisodes se terminant après une année d'observation). Ceci permet d'estimer un modèle de Weibull, dans lequel on a introduit une seule constante ; le programme s'écrit :

```
data cal;set cal;
if fin>=12;
y=fin-deb+1;logy=log(y);* Durée totale *;
if deb<12 then t=12-deb+1;else t=0;* Durée écoulée en T=12 *;
if t>0 then logt=log(t);else logt=0;
d=1-cens;

proc nlin data=cal sigsq=1 method=marquardt;
Parms a=0.5 b0=-1;* Valeurs initiales *;
_xb_=b0;
_lsurv_=(y**a)*exp(_xb_);* log de la survie *;
_l0_=(t**a)*exp(_xb_);* log de la survie en t *;
_loss_=- ( (_xb_+log(a)+(a-1)*logy)*d-_lsurv_+_l0_ ) ;
* opposé de la log vraisemblance *;
```

```
der.b0=(d-_lsurv_+_l0_);* dérivée par rapport à la constante *;  
der.a=( (1/a +logy)*d-_logy*_lsurv_+logt*_l0_);  
* dérivée par rapport au paramètre *;  
model y=y-1;* instruction générant un résidu de 1 *;  
run;
```

| Non-Linear Least Squares Iterative Phase | | Dependent Variable Y | | Method: Marquardt |
|--|------|----------------------|-----------|-------------------|
| Itérations ⇨ | Iter | A | B0 | Sum of Loss |
| | 0 | 0.500000 | -1.000000 | 12361.538896 |
| | 1 | 0.989082 | -2.740307 | 11761.739655 |
| | 2 | 0.837796 | -2.024140 | 11615.776045 |
| | 3 | 0.949467 | -2.359131 | 11606.890895 |
| | 4 | 0.879603 | -2.110524 | 11601.479401 |
| | 5 | 0.925816 | -2.266491 | 11600.097315 |
| | 6 | 0.894950 | -2.159333 | 11599.240141 |
| | 7 | 0.915325 | -2.229106 | 11598.939060 |
| | 8 | 0.901666 | -2.181949 | 11598.783346 |
| | 9 | 0.910730 | -2.213094 | 11598.720860 |
| | 10 | 0.904666 | -2.192193 | 11598.691115 |
| | 11 | 0.908702 | -2.206075 | 11598.678469 |
| | 12 | 0.906006 | -2.196789 | 11598.672670 |
| | 13 | 0.907803 | -2.202972 | 11598.670142 |
| | 14 | 0.906603 | -2.198843 | 11598.669001 |
| | 15 | 0.907403 | -2.201595 | 11598.668498 |
| | 16 | 0.906869 | -2.199758 | 11598.668273 |
| | 17 | 0.907225 | -2.200983 | 11598.668173 |

NOTE: Convergence criterion met.

| Non-Linear Least Squares Summary Statistics | | | Dependent Variable Y |
|---|------|----------------|------------------------------|
| Source | DF | Sum of Squares | Mean Square |
| Regression | 2 | 1042551.0000 | 521275.5000 ⇨ <i>Calculs</i> |
| Residual | 4941 | 4943.0000 | 1.0004 ⇨ <i>parasites</i> |
| Uncorrected Total | 4943 | 1047494.0000 | ⇨ |
| (Corrected Total) | 4942 | 490227.1062 | |
| Sum of Loss | | 11598.6682 | ⇨ - <i>Log vraisemblance</i> |

| Parameter | Estimate | Asymptotic Std. Error | Asymptotic 95 % Confidence Interval | |
|-----------------|--------------|-----------------------|-------------------------------------|---------------|
| | | | Lower | Upper |
| Estimateurs ⇨ A | 0.907225270 | 0.01514665370 | 0.8775305635 | 0.9369199764 |
| corrigés ⇨ B0 | -2.200983121 | 0.05002984674 | -2.2990656175 | -2.1029006239 |

Asymptotic Correlation Matrix

| Corr | A | B0 |
|--|--------------|--------------|
| ff | | |
| A | 1 | -0.949256936 |
| B0 | -0.949256936 | 1 |

Quelques références

- **BONNAL L., FOUGERE D.**, « Les déterminants individuels de la durée du chômage », *Economie et Prévision*, 1990-5, 96, p.45-82.
- **CARON N.**, « Les principales techniques de correction de la non-réponse et les modèles associés », *Document de travail Méthodologie Statistique*, 9604.
- **CASES C., LOLLIVIER S.**, « L'économétrie des modèles de durée avec SAS », *Document de travail CREST*, 9344BIS.
- **GOURIEROUX C.**, « Econométrie des variables qualitatives », *Economica*, 1989.
- **GOURIEROUX C., MONFORT A.**, « Modèles de durée et effets de génération », *Document de travail CREST*, 9125.
- **LANCASTER T.**, « The Econometric Analysis of Transition Data », *Econometric Society Monographs*, Cambridge University Press, 1990.
- **LOLLIVIER S., VERGER D.**, « D'une variable discrète à une variable continue : une application de la méthode des résidus simulés », in Mélanges en l'honneur de E.Malinvaud, *Economica*, 1988. *CREST*, 9125.
- **MOREAU A.**, « Econométrie des variables de durée », *Note Département Recherche*, n°123/G305, 1989
- **NICKELL S.**, « Estimating the Probability of Leaving Unemployment », *Econometrica*, 1987, 47, 1249-1266.
- **De TOLDI M., GOURIEROUX C., MONFORT A.**, « On Seasonal Effects in Duration Models », *Document de travail CREST*, 9216.
- **VISSER M.**, « Analysis of Labour Market Histories with Panel Data », *Document de travail CREST*, 9209.