L'économétrie des modèles de durée avec SAS. Présentation et mise en œuvre

C. Cases [1]
S. Lollivier [2]

1 Introduction

L'analyse économétrique des données de durée est une discipline assez récente. Les premiers manuels méthodologiques appliqués aux données économiques datent, en effet, du début des années 80, quand un développement suffisant aussi bien de la théorie des probabilités (processus), de l'analyse statistique et des moyens de calcul informatiques ont été atteints.

L'analyse des durées a d'abord été celle des durées de vie, et a été menée par les démographes et les actuaires. Elle est également très utile en biométrie et en statistique médicale, où elle sert à modéliser et à comparer des survies de malades suivant différents traitements, des durées de rémission... Un autre domaine d'application est traditionnellement celui des contrôle de fiabilité de matériels (taux de pannes de machines ou de systèmes). En économie, les domaines d'application privilégiés des modèles de durées sont les durées de chômage ou d'emploi des individus, mais ils peuvent être appliqués à des sujets très variés (durée de vie des entreprises, durée de remboursement d'un emprunt tenant compte des remboursements anticipés...).

Le présent document de travail vise à donner les éléments nécessaires à la modélisation de durées à l'aide du logiciel SAS. Il comporte d'abord une présentation synthétique des principaux outils probabilistes nécessaires, et des grandes catégories de modèles économétriques utilisés, ainsi que des éléments sur les méthodes d'estimation de ces modèles. Il détaille ensuite l'utilisation des diverses procédures SAS qui peuvent être utilisées pour réaliser ces estimations, en les illustrant d'exemples.

[1]CREST
[2]CREST

2 Caractériser la loi des variables 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 sur 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 est qu'elles peuvent s'interpréter facilement comme résultant d'un processus stochastique sous-jacent. Ce processus rend compte des dates de changements d'état d'un 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.

Présentons d'abord les trois fonctions les plus utilisées pour caractériser la loi d'une durée. Pour cela, on notera T la variable de durée, $f(t)$ et $F(t)$ sa densité de probabilité et sa fonction de répartition.

On appelle **fonction de survie** $S(t)$ la probabilité que la durée soit plus grande que t, soit

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

On appelle **fonction de hasard** $h(t)$ la probabilité que la durée soit comprise entre t et $t + dt$, sachant qu'elle est plus grande que t, soit

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

$h(t)$ représente le taux instantané de sortie de l'état que l'on observe. Si, par exemple, on mesure des durées de chômage, $h(t)$ représentera le taux de sortie de chômage à la date t, c'est -à-dire la probabilité de sortir du chômage dans un très petit intervalle de temps après t, sachant que l'on était chômeur en t. Si

l'on s'intéresse à la durée de vie des individus, $h(t)$ sera un risque de mortalité à un âge donné.

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 le cas du dernier exemple.

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,

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

d'où

$$S(t) = \exp(-\int_0^t h(u)\, du).$$

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 observées 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).

3 Les lois de probabilité de base

La loi de référence pour les modèles de durée est la **loi exponentielle**. Elle a une propriété importante : elle est la seule à avoir un hasard constant. La valeur de ce hasard est le seul paramètre de la loi. 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 du modèle exponentiel qu'il est "sans mémoire"[1]. La valeur du hasard est le seul paramètre de la loi. Ses caractéristiques sont les suivantes :

$h(t) = \theta$
$S(t) = \exp(-\theta t)$
$f(t) = \theta \exp(-\theta t)$
$r(t) = 1/\theta.$

La **loi de Weibull** généralise la loi exponentielle, puisque la durée Y est supposée telle que Y^α suit une loi exponentielle de paramètre θ. C'est donc une loi à deux paramètres α, θ telle que :
$h(t) = \alpha\theta t^{\alpha-1}$
$S(t) = \exp(-\theta t^\alpha)$
$f(t) = \alpha\theta t^{\alpha-1} \exp(-\theta t^\alpha)$.
Le hasard de la loi de Weibull est monotone, croissant si $\alpha > 1$ et décroissant si $\alpha < 1$. De plus, la loi de Weibull englobe la loi exponentielle pour $\alpha = 1$.

La **loi log-normale** et la **loi log-logistique** permettent de représenter des hasards avec un mode (croissants, puis décroissants). La durée T sera alors telle que $\frac{\log T - m}{\sigma}$ suit respectivement une loi normale $N(0,1)$ ou une loi logistique. Le hasard de la loi log-normale a une expression analytique inconnue, qui dépend du ratio de Mills :

$$h(t) = \frac{\phi(\frac{\log t - m}{\sigma})}{t\sigma(1 - \Phi(\frac{\log t - m}{\sigma}))},$$

où ϕ et Φ sont la densité et la fonction de répartition (calculable numériquement seulement) de la loi normale centrée réduite.

Pour éviter de manipuler une forme aussi complexe, on préfère le plus souvent utiliser la loi log-logistique, qui est très proche de la loi log-normale, et dont le hasard s'écrit :

[1] Le processus sous-jacent est markovien.

$$h(t) = \frac{\theta t^{(1/\sigma)-1}}{\sigma(\theta t^{1/\sigma} + 1)},$$

où $\theta = \exp(-m/\sigma)$. Pour $\sigma < 1$, le hasard présente un mode ; pour $\sigma \geq 1$, il est monotone décroissant, avec ou sans asymptote en 0.

On peut construire d'autres familles de lois. Pour plus de détails, se référer à la bibliographie en fin de volume.

4 Les grands principes de l'économétrie des durées

4.1 Modèle structurel, modèle réduit

Pour estimer un modèle de durée, la méthode la plus simple est d'observer des durées et de procéder directement à l'estimation des paramètres de la loi de probabilité de la variable aléatoire, par exemple sa fonction de hasard. Mais cette fonction s'interprète le plus souvent comme résultant d'un comportement particulier. Ce sont, en fait, les caractéristiques de ce comportement qui servent, en dernier ressort, à comprendre la distribution des durées étudiées. On peut donc aussi chercher à modéliser directement ces comportements. Dans le premier cas, on dit que l'on estime la forme réduite du modèle. Dans le second, on en analyse la forme structurelle. Ce sont des estimations de formes réduites dont traitera ce fascicule.

Pour illuster la différence entre un modèle structurel et un modèle réduit, prenons l'exemple classique de l'analyse des durées de chômage à l'aide d'un modèle de recherche d'emploi. On suppose qu'un individu au chômage reçoit des offres d'emploi à chaque moment avec une probabilité constante λ. Ces offres sont caractérisées par leur salaire w qui est tiré aléatoirement dans une distribution de fonction de répartition F, connue à l'avance par le chômeur. A chaque date, l'individu reçoit une indemnité b s'il est au chômage. Il peut refuser ou accepter une offre, mais ne revient jamais sur une décision passée. On suppose que sa stratégie consiste à maximiser son espérance de revenu sur une durée de vie infinie. Une fois accepté, l'emploi est définitif et le salaire ne change plus. On montre alors que la stratégie optimale du chômeur est d'accepter une offre seulement si son salaire dépasse un montant minimum ξ appelé salaire de réserve, qui est une fonction assez complexe de tous les paramètres λ, b, F et de son taux d'actualisation[2]. La fonction de hasard s'écrit alors $h(t) = \lambda(1-F(\xi))$. Dans ce cas simple, elle ne dépend pas de t : le modèle est dit stationnaire. Si λ ou b varie avec t, ou si la durée de vie est finie, le salaire de réserve et le hasard dépendront de t. Un modèle structurel estimera séparément λ, b, F. Un modèle réduit essaiera d'estimer globalement la fonction de hasard. La forme du hasard et son sens de variation avec t est une des questions fondamentales en économétrie des durées.

[2] $\xi = b + \frac{\lambda}{\rho}\int_{\xi}^{\infty}(w-\xi)\,dF(w)$

4.2 Modèle paramétrique, non paramétrique, semi-paramétrique

Dans l'exemple précédent, plusieurs stratégies sont possibles pour l'estimation directe de la fonction de hasard. On peut supposer que la variable de durée suit une loi de probabilité donnée, par exemple une loi exponentielle, une loi de Weibull...On peut alors écrire la vraisemblance de l'échantillon observé, et estimer ses paramètres par maximisation. Le modèle est alors dit **paramétrique**. On peut aussi introduire dans le modèle des variables exogènes qui déterminent la valeur de certains paramètres (voir ci-dessous). Des exemples d'écriture de vraisemblance seront traités dans la section 6.

Certaines méthodes permettent de s'affranchir d'une spécification particulière de la loi des durées. En effet, celles-ci peuvent être trop contraignantes (difficulté de modéliser un hasard à plusieurs modes, par exemple), ou trop peu robustes (les résultats peuvent être très différents selon la spécification choisie). Selon que l'on laisse libre l'ensemble ou une partie de la spécification de la loi de la durée, on parlera de modèles **semi-paramétriques ou non paramétriques**. Ils sont cependant parfois plus difficiles à programmer et nécessitent souvent plus de données. Cependant, certains modèles courants peuvent être traités très simplement avec des procédures SAS. Il s'agit en particulier de l'estimateur non paramétrique le plus courant, dit de **Kaplan-Meier** (PROC LIFETEST), et du modèle semi-paramétrique de **Cox** (PROC PHREG), dont les grandes caractéristiques seront décrites plus loin.

4.3 Introduction de variables exogènes

L'estimation des fonctions de hasard doit a priori s'effectuer sur des populations homogènes. Si la population regroupe des catégories dont les lois de durées sont différentes, le risque est en effet de conclure faussement à une décroissance de la fonction de hasard. Le mécanisme qui mène à ce biais est connu sous le nom de "mover-stayer" : supposons un mélange à part égales de deux populations à hasards (ou risques) constants, mais différents. Au fil du temps, les individus de la population de risque le plus élevé sortant plus vite de l'état observé, la population des survivants comportera de plus en plus d'individus à risque faible, et les sorties seront ainsi de moins en moins fréquentes.

Pour éviter ce risque de mauvaise interprétation, il est possible de partager l'échantillon observé en sous-échantillons (ou **strates**) les plus homogènes possibles. Par exemple, on peut envisager d'étudier séparément les durées de chômage selon le sexe, le diplôme et la classe d'âge. Procéder ainsi suppose qu'il reste dans chaque sous-échantillon suffisamment d'individus pour que l'estimateur conserve de bonnes propriétés asymptotiques. On peut aussi spécifier une forme

paramétrique particulière dans laquelle les paramètres s'expriment en fonction de **variables** exogènes.

Il existe plusieurs catégories de familles paramétriques qui permettent de procéder ainsi. Les plus courantes sont les familles à hasard proportionnel et les familles à hasard accéléré.

Dans **les familles à hasard proportionnel**, la fonction de hasard a pour forme générale :

$$h(t) = h_0(t)\phi(X, \beta).$$

$h_0(t)$ est appelé "hasard de base", et $\phi(X, \beta)$ est une fonction positive des exogènes X, β étant un vecteur de paramètres. On choisit en général $\phi(X, \beta) = \exp(X\beta)$. Le nom de cette famille de lois tient à ce que des valeurs différentes des variables exogènes aboutissent à des valeurs proportionnelles du hasard. En particulier, si le hasard de base présente un mode, ce sera le même pour tous les individus, ce qui peut être très restrictif. Le hasard de base peut être estimé par la méthode du maximum de vraisemblance en spécifiant une forme paramétrique particulière, ou bien par une méthode non paramétrique (on parle alors d'une estimation semi-paramétrique pour h, voir plus loin le détail d'une méthode : modèle de Cox).

Dans les **familles à hasard accéléré**, la fonction de hasard a pour forme générale :

$$h(t, X, \beta) = h_0[t \exp(X\beta)] \exp(X\beta).$$

Les variables exogènes ont alors un effet de paramètre d'échelle sur les durées : tout se passe comme si la durée T d'un individu de la "catégorie" X s'écrivait $T_0 \exp(-X\beta)$, où T_0 serait la durée de vie de la catégorie de référence. Tout se passe donc comme si le temps avançait plus ou moins rapidement pour les différents types d'individus. Cette écriture permet d'écrire simplement les modèles à durée de vie accélérée sous la forme :

$$\log T = -X\beta + \log T_0.$$

Cette écriture peut faire penser à un modèle de régression linéaire, où $\log T_0$ jouerait le rôle de la perturbation. Le problème principal est que dans le cas général, cette "perturbation" n'est pas d'espérance nulle, et que les moindres carrés ordinaires ne peuvent être appliqués pour estimer β que dans des cas très particuliers (pas de données censurées). Dans la plupart des cas, on doit spécifier la loi de $\log T_0$ et estimer par le maximum de vraisemblance. C'est la méthode utilisée par la procédure LIFEREG de SAS (voir plus loin). Il existe des méthodes semi-paramétriques qui évitent de spécifier une loi pour $\log T_0$, mais elles ne sont pas disponibles sous formes de procédures SAS.

5 Problèmes particuliers

5.1 Les données censurées

Une des particularités les plus fréquentes des données 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**. Le type de censure le plus fréquent est ainsi la "censure à droite". Supposons que l'on observe toutes les personnes entrant au chômage entre deux dates T_1 et T_2. Pour les personnes ayant retrouvé un emploi en T_2, la durée sera parfaitement observée. Pour les personnes toujours au chômage en T_2, on sait seulement que la durée de chômage est supérieure à ce que l'on a observé (on parle alors d'une ancienneté de chômage). Si l'on ne tient pas compte de ce phénomène, la loi de durée que l'on estimera sera biaisée et conduira à des espérances de durée plus courtes que la réalité.

Il existe différents types de censure qui ne seront pas tous détaillés ici. Il est en général assez simple de tenir compte de la censure si elle intervient de manière indépendante du mécanisme de sortie, c'est-à-dire si la loi des durées censurées est bien la même que celle des durées non censurées. Les procédures SAS présentées par la suite traitent toujours ces cas simples.

5.2 Les fichiers de stock

Un cas, lui aussi fréquent, mais non pris en compte dans les procédures SAS est celui des échantillons construits à partir de fichiers de stock. Dans le cas des durées de chômage, par exemple, il est fréquent d'observer les durées d'individus se trouvant au chômage à la date de début de l'enquête (par exemple en tirant l'échantillon dans un fichier ANPE). Cette méthode de tirage de l'échantillon introduit un biais sur les durées observées, appelé **biais de sélection endogène** (stock sampling). Pour bien s'en persuader, il suffit de raisonner sur l'ensemble des personnes entrées au chômage à une même date $-e$ (on posera que la date de tirage de l'échantillon vaut 0). Parmi cette "cohorte", seuls figureront dans les fichiers de chômeurs en 0 les individus dont la durée de chômage est plus grande que e. Les autres auront quitté les fichiers auparavant. Pour chaque cohorte d'entrants, la probabilité de figurer dans l'échantillon sera nulle pour les durées les plus courtes. Ce mode de sélection particulier conduit donc à surestimer les durées moyennes si l'on ne corrige pas de ce biais. **Dans la pratique, cette surestimation peut être très importante, et conduire à multiplier par 2 ou 3 les espérances de durée.** Il existe des méthodes de correction, qui font

souvent des hypothèses fortes sur la stabilité des lois de durée pour l'ensemble des cohortes d'entrants. La plus simple à mettre en oeuvre est l'estimation par le maximum de vraisemblance conditionnel. Elle n'est cependant pas incluse dans les procédures SAS et nécessite d'utiliser ou de programmer complètement un algorithme de maximisation. La PROC NLIN peut être utilisée dans ce cas, après avoir calculé formellement la vraisemblance et le score (vecteur des dérivées par rapport aux paramètres).

5.3 L'hétérogénéité non observée

On a vu précédemment qu'il était nécessaire de procéder à des estimations sur des populations homogènes ou bien d'inclure des variables exogènes dans la spécification des lois de durées. Le problème de l'hétérogénéité reste entier lorsqu'elle résulte de variables omises ou d'un caractère non observable des individus. Dans ce cas, on peut conclure faussement à une décroissance du hasard avec t, et même obtenir des estimateurs biaisés pour les coefficients des variables exogènes incluses dans le modèle. Pour remédier à ce problème, on introduit généralement un facteur d'hétérogénéité multiplicatif v dans la fonction de hasard, pour lequel on spécifie une loi particulière de probabilité, discrète ou continue. La vraisemblance du modèle peut alors être écrite en intégrant sur la loi de v, dont on estime les paramètres (et éventuellement le support, s'il s'agit d'une loi discrète) comme les autres éléments du modèle. Ce type de modèle, qui devient assez courant en pratique, n'est pas non plus disponible en standard dans SAS.

5.4 Les exogènes variant dans le temps

L'introduction d'exogènes dans le modèle n'a été envisagée que dans le cas où elles mesurent des caractéristiques constantes au cours de la durée d'observation. Il est possible d'étendre le modèle au cas d'une variable $x(t)$. Pour cela, il faut créer autant de variables que de sous-périodes pendant lesquelles x est constante, c'est-à-dire conditionner le modèle par l'ensemble des valeurs de la variable. Cela ne pose pas de problème dans le principe, mais le modèle peut devenir difficile à estimer si l'on introduit plusieurs variables de ce type, le nombre de coefficients à estimer pouvant alors devenir très important.

6 Estimation et tests dans les modèles paramétriques

L'une des difficultés d'estimation des modèles de durées est l'impossibilité d'appliquer les modèles de régression habituels, sauf dans des cas très particuliers. On a vu précédemment que l'on pouvait penser à écrire un modèle de la forme :

$$\log T = X\beta + U,$$

où U est une perturbation. Mais les moindres carrés ordinaires ne sont généralement pas convergents, sauf dans le cas où les données observées ne sont pas censurées. La méthode utilisée est donc presque toujours le maximum de vraisemblance.

6.1 Ecriture de la vraisemblance dans les modèles de durée

Supposons que, dans le cas d'un échantillon de taille N, soient observées des durées, complètes ou censurées, t_i pour chaque individu $i = 1, ...N$. Cela revient à disposer, en plus de la valeur de t_i, d'une variable indicatrice de censure C_i, telle que $c_i = 1$ si la durée t_i est censurée, et 0 sinon.

La vraisemblance du modèle s'écrit alors :

$$L = \prod_{i=1}^{n} f(t_i)^{c_i} S(t_i)^{(1-c_i)}.$$

En effet, la probabilité qu'une durée soit censurée en t_i, donc supérieure où égale à t_i est la valeur de la survie $S(t_i)$.

La log-vraisemblance a donc pour forme

$$\log L = \sum_{i=1}^{n} c_i \log f(t_i) + \sum_{i=1}^{n} (1 - c_i) \log S(t_i).$$

Cette expression peut se simplifier en utilisant la relation $h(t_i) = f(t_i)/S(t_i)$, ce qui donne

$$\log L = \sum_{i=1}^{n} c_i \log h(t_i) + \sum_{i=1}^{n} \log S(t_i).$$

Lorsque l'on spécifie une forme particulière pour h et donc pour S, avec éventuellement introduction de variables exogènes, on obtient simplement la valeur de la fonction à maximiser en calculant $\log h(t_i)$ et $\log S(t_i)$.

6.2 Algorithmes de maximisation

Les procédures SAS utilisent des algorithmes de résolution numérique pour maximiser la log-vraisemblance. Le plus utilisé est l'algorithme de **Newton-Raphson**, dont le principe est rappelé ci-dessous. La procédure NLIN permet, de plus, de choisir entre différents algorithmes de calcul, plus ou moins précis ou rapides selon les cas (voir sections 9 et 10).

Pour maximiser la log-vraisemblance $\log L = l(\beta)$, il faut chercher une solution β^* qui annule la dérivée de l par rapport à β, également appelée vecteur du score. Pour cela, l'algorithme de Newton-Raphson part d'une valeur initiale $\hat{\beta}_0$, et résoud itérativement l'équation :

$$\hat{\beta}^{j+1} = \hat{\beta}^j - \left[\frac{\partial^2 l(\hat{\beta}^j)}{\partial\beta\partial\beta'}\right]^{-1} \frac{\partial l(\hat{\beta}^j)}{\partial\beta}.$$

Le second terme du membre de droite est appelé le pas de l'algorithme. Les itérations se poursuivent jusqu'à ce que $l(\hat{\beta}^{j+1}) - l(\hat{\beta}^j)$ soient très proches (par défaut, en général, dans les procédures SAS, l'écart entre les deux fonctions doit être finalement inférieur à 10^{-4}). De plus, il est vérifié à chaque étape que $l(\hat{\beta}^{j+1}) - l(\hat{\beta}^j) > 0$. Si ce n'est pas le cas, on recalcule un nouveau $\hat{\beta}^{j+1}$ en réduisant le pas.

Il faut noter que, dans les cas où la vraisemblance n'est pas strictement concave, **on n'est nullement assuré de la convergence de l'algorithme vers le maximum cherché**, puisqu'aucune condition de second ordre n'a été vérifiée dans un premier temps. Si l'on a des doutes, il est possible de fixer des valeurs initiales assez éloignées et d'effectuer plusieurs fois la maximisation afin de comparer les résultats. D'autre part, il est évident que la convergence est beaucoup plus rapide si les valeurs initiales sont bien choisies, par exemple si elles sont le résultat d'une procédure d'estimation moins affinée.

6.3 Propriétés de l'estimateur

De manière générale, l'estimateur du maximum de vraisemblance est **asymtotiquement convergent et normal**, de variance asymptotique estimée:

$$\hat{V}(\hat{\beta}) = -\left[\frac{\partial^2 l(\hat{\beta}^j)}{\partial\beta\partial\beta'}\right]^{-1}.$$

La connaissance de cette loi asymptotique est essentielle pour effectuer des tests de spécification, comme nous le verrons ci-dessous.

6.4 Cas particulier d'un modèle de Weibull

Dans le cas d'un modèle de durée simple (sans sélection endogène), la log-vraisemblance s'écrit donc :

$$\log L = \sum_{i=1}^{n} c_i \log h(t_i) + \sum_{i=1}^{n} \log S(t_i),$$

où C_i est la variable indicatrice de censure. Dans le cas d'un modèle de Weibull à hasard proportionnel, le hasard s'écrit :

$$h(t_i) = \alpha(\exp(x_i'\beta))t_i^{\alpha-1},$$

où x_i' est le vecteur ligne des valeurs prises par les variables exogènes pour l'individu i. La survie a pour forme :

$$S(t_i) = \exp[-\exp(x_i'\beta)t_i^{\alpha}].$$

La log-vraisemblance vaut donc :

$$\log L = \sum_{i=1}^{n} c_i[\log\alpha + x_i'\beta + (\alpha-1)\log t_i] - \sum_{i=1}^{n}(\exp(x_i'\beta))t_i^{\alpha}].$$

Les dérivées partielles de la log-vraisemblance par rapport à α et β valent :

$$\frac{\partial \log L}{\partial \alpha} = \sum_{i=1}^{n} c_i[\frac{1}{\alpha} + \log t_i] - \sum_{i=1}^{n} \exp(x_i'\beta)t_i^{\alpha}\log t_i .$$

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} c_i x_i' - \sum_{i=1}^{n} x_i' \exp(x_i'\beta)t_i^{\alpha}.$$

6.5 Tests sur les paramètres

L'estimateur du maximum de vraisemblance est, on l'a vu, asymptotiquement normal. Cette propriété va permettre d'effectuer des test asymptotiques[3] sur les paramètres estimés. Le test le plus immédiat porte sur la significativité d'une variable exogène. On peut aussi devoir tester l'existence d'une contrainte linéaire sur les paramètres : dans l'exemple du modèle de Weibull ci-dessus, tester si $\alpha = 1$ revient à évaluer la significativité d'un modèle exponentiel. On peut également se demander si les coefficients de deux caractéristiques, par exemple des variables représentant des tranches d'âges voisines, sont différents ou semblables, etc...

[3] C'est-à-dire quand le nombre d'individus étudiés est grand.

Il existe **trois grands types de tests asymptotiques** applicables dans ce cas. Présentons-les dans le cas général du test d'une contrainte linéaire sur les paramètres. On veut tester :

H_0 : $L\beta = c$, contre

H_1 : $L\beta \neq c$, avec L et c, matrices de coefficients réels de taille convenable.

L'idée de ces tests est simple. Le premier (**test du rapport de vraisemblance**) compare la valeur des log-vraisemblances sous les deux hypothèses. Si elles sont assez proches, on pourra accepter H_0. Cela nécessite de pouvoir calculer simplement les estimateurs β_0 et β_1 de β sous les deux hypothèses. On montre que la quantité :

$LR = 2[\log L(\beta_1) - \log L(\beta_0)]$ converge en loi vers un $\chi^2(r)$, où r est le rang de L, c'est-à-dire le nombre de contraintes indépendantes sur les β_j.

L'hypothèse nulle sera rejetée si la valeur calculée de LR dépasse un seuil critique.

Le second (**test de Wald**) revient à évaluer la contrainte à l'aide de β_1. Si la valeur trouvée est assez proche de 0, on peut accepter l'hypothèses nulle. On montre que la quantité

$W = (L\beta_1 - c)[L\hat{V}(\beta_1)L'](L\beta_1 - c)$ converge en loi vers un $\chi^2(r)$, où r est le rang de L.

L'hypothèse nulle sera rejetée si la valeur calculée de W dépasse un seuil critique. Ce test est particulièrement utilisé quand l'estimation sous l'hypothèse alternative est plus simple que sous l'hypothèse nulle.

Le troisième (**test du score**) revient à calculer le vecteur du score sous l'hypothèse nulle. Si la valeur trouvée pour $L\beta_1 - c$ est assez proche de 0, on peut penser que l'on ne s'éloigne pas trop du maximum de vraisemblance et que l'on peut donc accepter l'hypothèse nulle. On montre que la quantité :

$$S = \left(\frac{\partial L(\beta_0)}{\partial \beta}\right)' \left[\frac{\partial^2 L(\beta_0)}{\partial \beta \partial \beta'}\right]^{-1} \left(\frac{\partial L(\beta_0)}{\partial \beta}\right)$$

converge en loi vers un $\chi^2(r)$, où r est le rang de L.

L'hypothèse nulle sera rejetée si la valeur calculée de S dépasse un seuil critique. Ce test est particulièrement utilisé quand l'estimation du score sous l'hypothèse nulle est simple.

Ces tests sont asyptotiquement équivalents, c'est-à-dire que pour N infini, ils amèneront à prendre les mêmes décisions. Ils sont fréquemment proposés dans les procédures SAS qui seront développées par la suite. Le test du rapport

de vraisemblance peut être calculé simplement par l'utilisateur même s'il n'est pas directement calculé en standard [4].

Notons également que si l'on veut simplement tester la nullité d'un seul coefficient du modèle, il est plus simple d'utiliser un test de Student qui est parfaitement équivalent au test de Wald. La statistique de Student s'écrit, comme dans un modèle de régression habituel,

$$T = \frac{\hat{\beta}_j}{\sqrt{\hat{V}(\hat{\beta}_j)}}.$$

C'est la racine carrée de la statistique de Wald calculée dans le cas précis où la contrainte est $\beta_j = 0$. T suit asymptotiquement une loi normale centrée réduite. Il suffit alors de comparer T au seuil de significativité habituel de la loi normale (environ 2 pour un risque de 5%). Si T est inférieur au seuil, on acceptera l'hypothèse nulle.

[4] Il suffit en effet de procéder à deux estimations, l'une contrainte et l'autre non, et de calculer LR à l'aide de la valeur de la log-vraisemblance au maximum, qui figure dans toutes les éditions de résultats des procédures.

7 Un estimateur non paramétrique : Kaplan-Meier

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:

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

avec n_t : nombre de durées inférieures à t et N : nombre total d'observations. Dans SAS, cette fonction de répartition empirique est simplement donnée 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 $t_0, t_1, ...t_N$, 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 sera :

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

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

$$E(T) = \int_0^\infty uf(u)\,du = \int_0^\infty S(u)\,du,$$

on peut utiliser l'estimateur suivant :

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

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 :

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

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 $(T_i, D-i)$)). Il suffit alors de comparer à 1,96 le rapport $U/\sqrt{V_0(U)}$.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, un vecteur Υ de statistiques de rangs dont les composantes sont définies par $\Upsilon_k = \sum_i \sum_{j \neq i} U_{ij} Z_{ik}$ où Z_{ik} est une variable indicatrice d'appartenance à la strate k.

La statistique globale utilisée pour le premier type d'hypothèse est $\Upsilon' V^- \Upsilon$ (où V^- est une inverse généralisée de la variance estimée de Υ) 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.

8 Une estimation semi-paramétrique : le modèle de Cox

Une méthode d'estimation semi-paramétrique est disponible dans la PROC PHREG de SAS(Versions 6). Elle concerne les modèles à hasard proportionnels présentés dans la partie 4.3 avec la spécification suivante pour la fonction de hasard:

$$h(t/x;\beta) = \exp(x\beta)h_0(t),$$

où h_0 est le hasard de base. Elle repose sur la maximisation de la "vraisemblance partielle" de Cox.

8.1 Vraisemblance partielle de Cox

Reprenons le cas où l' on a ordonné les valeurs des I durées différentes observées: $t_1 < t_2 < ... < t_I$ et où il n'y a pas de censure. Soit comme précédemment $r(t_i)$ l' ensemble à risque en t_i.La probabilité pou que ce soit l' individu j de $r(t_i)$ qui sorte en t_i vaut:

$$\frac{h(t_i/x_j;\beta)}{\sum_{k\in r(t_i)} h(t_i/x_k;\beta)}$$

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 $h_0(t)$ figure dans de dénominateur et le numérateur, et elle vaut finalement:

$$\frac{\exp(x_j\beta)}{\sum_{k\in r(t_i)} \exp(x_k\beta)}$$

La vraisemblance partielle de Cox est le produit de ces probabilités pour l' ensemble des sorties (on supposera qu'il y en a en tout $S \leq N$):

$$L(\beta) = \prod_{i=1}^{S} \frac{\exp(x_j\beta)}{\sum_{k\in r(t_i)} \exp(x_k\beta)}$$

S' il n'y a pas de censure, elle s'interprètre comme la vraisemblance de la statistique de rang associée aux durées. L' estimateur semi-paramétrique de β va être obtenu en maximisant la log-vraisemblance partielle par rapport à β au moyen d'une méthode itérative(voir partie 6).

L'estimateur obtenu converge presque sûrement vers β et est asympotiquement normal.

8.2 Estimation non paramétrique du hasard de base

On préfère, en général, estimer directement la fonction de survie. Dans le modèle de Cox, cette fonction a une forme simple : elle s'écrit

$$S(t) = [S_0(t)]^{\exp X\beta}.$$

Cette relation découle de la définition du modèle et de la relation générale entre hasard et survie.

Kabfleish et Prentice en déduisent une méthode d'estimation de la "survie de base" en deux étapes. Dans une première étape, on estime β par une maximisation de vraisemblance partielle. Ensuite, β 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

$$\hat{\alpha}_i \exp(z_i'\beta) = 1 - \frac{\exp(z_i'\hat{\beta})}{\sum_{k\in r(t_i)} \exp(z_k'\hat{\beta})}$$

L' estimateur utilisé dans PHREG est celui de Breslow:

$$\hat{S}_0(t) = \prod_{t_i<t} \left[1 - \frac{d_i}{\sum_{k\in r(t_i)} \exp(z'k\hat{\beta})}\right]$$

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 $\hat{\beta}$ 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;

$$h(t) = h_0(t) \exp \beta.$$

La relation sur le hasard intégré s'écrira alors:

$$H(t) = H_0(t) \exp \beta,$$

d'où $\log H(t) - \log H_0(t) = \beta$. 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*).

9 Les procédures SAS d'analyse des durées

9.1 Estimations non paramétriques; la procédure LIFETEST

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.

Mise en oeuvre simplifiée (principales options).

```
PROC LIFETEST  < Options 1 >  ;  }
TIME variable  < Options 2 >  ;  } Instructions obligatoires

By variables                     }
ID variables                     }
STRATA variable  < options 3 >  ;  } Instructions facultatives
TEST variables                   }
```

Options 1 :

. DATA = ; précise la table SAS contenant les données.

. INTERVALS = *value* ; 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, *intervals = 5, 10 to 30 by 10* produit le découpage $[0,5), [5,10), [10,20), [20,30), [30,\infty)$.

. METHOD = *type* ; 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 (nécessaire sur les fichiers de données individuelles).

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

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

. OUTEST = data ; crée un fichier SAS contenant différents estimateurs pour chacun des intervalles des différents strates :

- variables BY et STRATA

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

. OUTEST = Data ; 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 ; elle peut 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. 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:

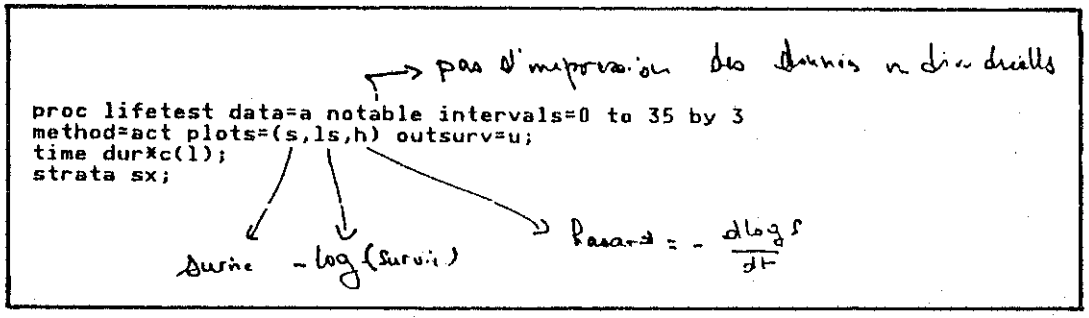
L'instruction TEST fournit une liste de covariables numériques dont on veut tester les liens avec la durée de vie.

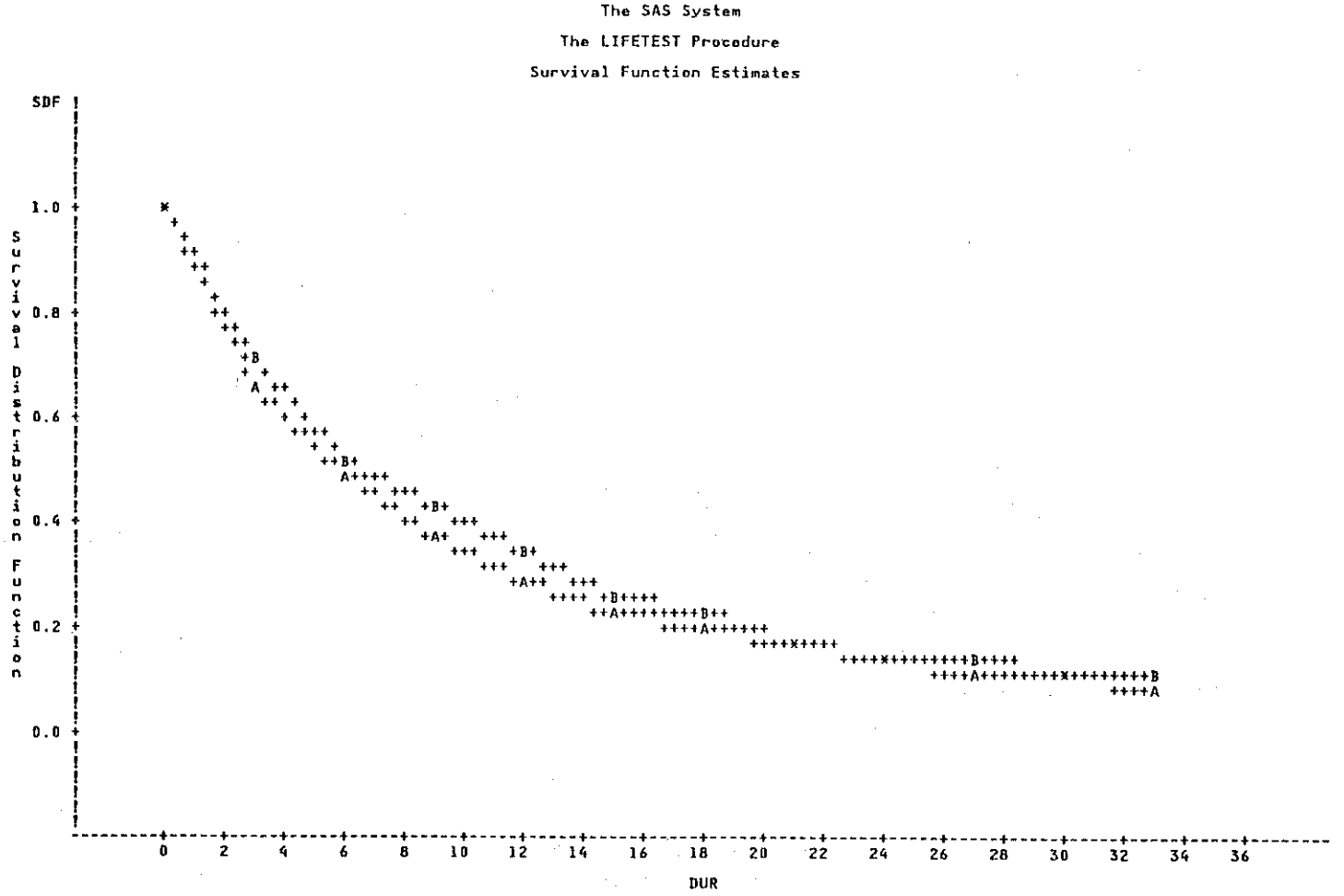
The SAS System

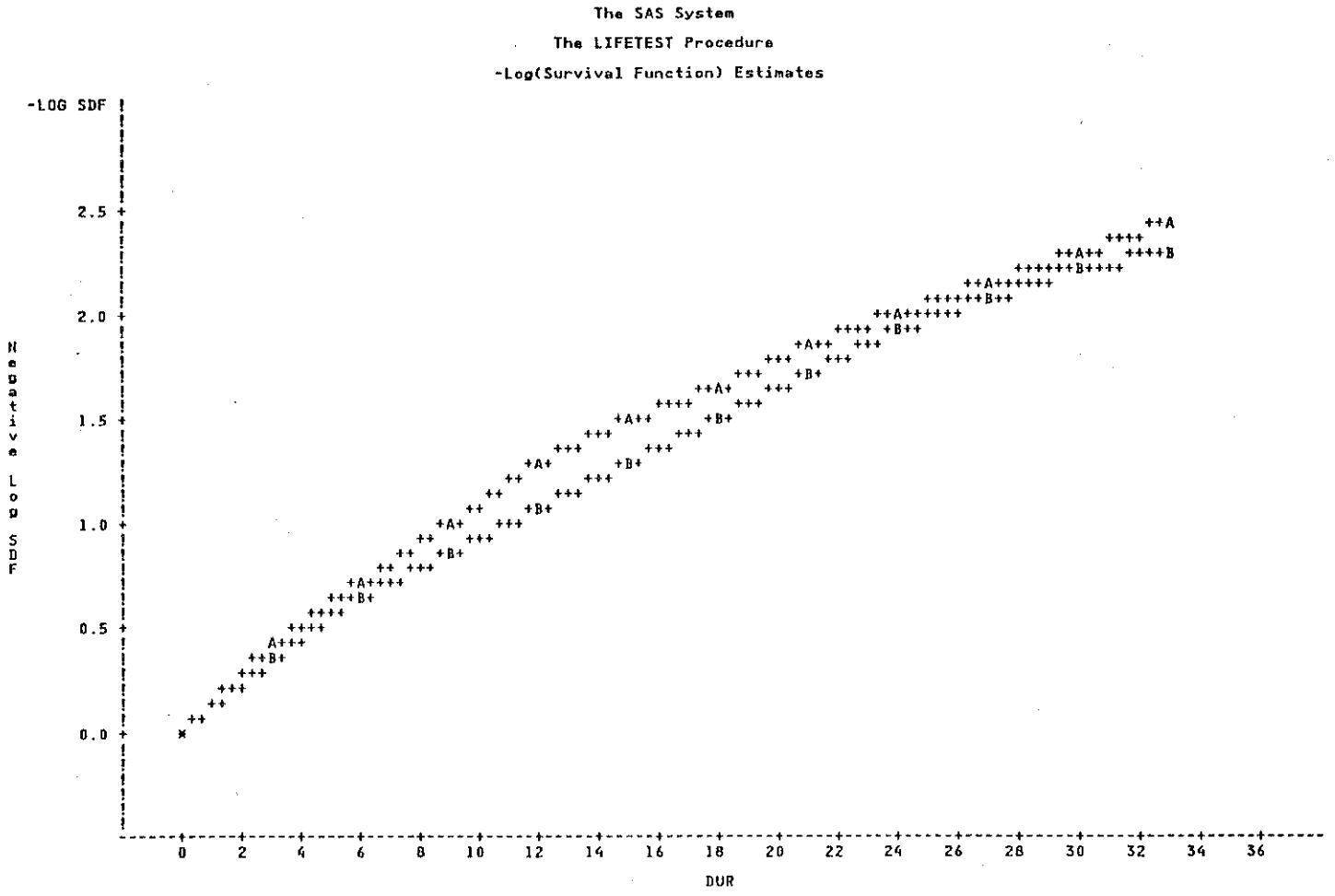
The LIFETEST Procedure

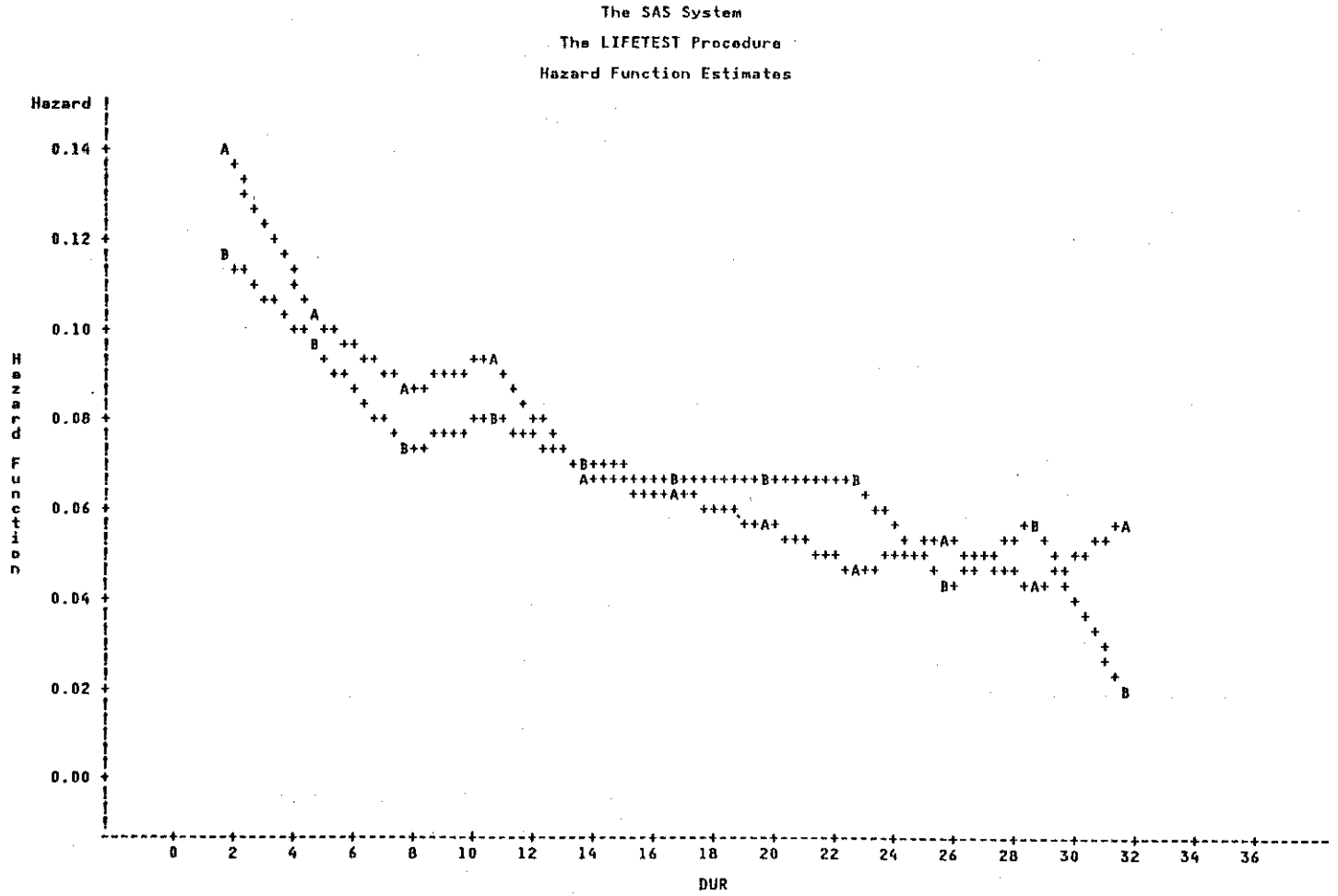
Summary of the Number of Censored and Uncensored Values

SX	Total	Failed	Censored	%Censored
1	11652	6548	5104	43.8036
2	13394	7150	6244	46.6179
Total	25046	13698	11348	45.3086




The SAS System
The LIFETEST Procedure
Survival Function Estimates
SDF
Survival Distribution Function
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
DUR


The SAS System
The LIFETEST Procedure
-Log(Survival Function) Estimates
-LOG SDF
Negative Log SDF
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
DUR


The SAS System
The LIFETEST Procedure
Hazard Function Estimates
Hazard
Hazard Function
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
DUR

The SAS System

The LIFETEST Procedure

Testing Homogeneity of Survival Curves over Strata
Time Variable DUR

Rank Statistics

SX	Log-Rank	Wilcoxon
1	490.98	8902672
2	-490.98	-8902672

Statistiques de rang par sous-groupes
Les deux statistiques sont opposées car il n'y a que deux sous-groupes.

Covariance Matrix for the Log-Rank Statistics

SX	1	2
1	3023.22	-3023.22
2	-3023.22	3023.22

Variance estimée du premier vecteur de statistiques de rang (Log Rank)

Covariance Matrix for the Wilcoxon Statistics

SX	1	2
1	8.698E11	-8.7E11
2	-8.7E11	8.698E11

Variance estimée du vecteur de statistiques de Wilcoxon

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	79.7357	1	0.0001
Wilcoxon	91.1198	1	0.0001
-2Log(LR)	114.0473	1	0.0001

Valeur test pour les deux tests de rangs

Probabilités faibles => rejet de l'hypothèse d'homogénéité des lois de durée entre hommes et femmes.

Test de rapport de vraisemblance

H_0 : les lois sont exponentielles de même paramètre dans les deux sous-groupes.

H_1 : paramètres différents (mais exponentielles)

hypothèses plus restrictives que les deux tests précédents.

9.2 Estimations paramétriques

9.2.1 la procédure LIFEREG.

Cette procédure estime des modèles à durée de vie accélérée [5]sous la forme :

$$y = xb + \sigma u$$

où $\exp(U)$ suit une loi connue (exponentielle, logistique, normale).

Elle fournit en sortie des estimateurs de b et σ.

LIEN AVEC DES MODÈLES CONNUS

Soit T la variable aléatoire représentant la durée de vie.

- dans le **modèle exponentiel**, $\log(\theta T) = U$ où $\exp(U)$ suit une loi exponentielle d'espérance 1. De ce fait,

$$y = \log(T) = -\log(\theta) + U.$$

Si on pose $\theta = \exp(x\beta)$, on obtient $\hat{\beta} = -\hat{b}$, en contraignant $\sigma = 1$.

- dans le **modèle de Weibull**, $\log(\lambda T) = \frac{U}{\alpha}$ où $\exp(U)$ suit à nouveau une loi exponentielle d'espérance 1.

La fonction de hasard s'écrit alors : $\theta(t) = \alpha\lambda^{\alpha}t^{\alpha-1}$.

Si on pose $\lambda^{\alpha} = \exp(x\beta)$ pour ramener à la spécification habituelle, on obtient :

$$y = \log(T) = -x\frac{\beta}{\alpha} + \frac{U}{\alpha}$$

et par conséquent : $\hat{\alpha} = \frac{1}{\hat{\sigma}}$ et $\hat{\beta} = -\frac{\hat{b}}{\hat{\sigma}}$

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

$$\bar{F}(t) = \frac{1}{1 + \exp(\exp(x\beta)t^{\alpha})}$$

On retrouve $\hat{\alpha} = \frac{1}{\hat{\sigma}}$ et $\hat{\beta} = -\frac{\hat{b}}{\hat{\sigma}}$

Pour la fonction **log-normale** :

[5]Les modèles exponentiels et de Weibull sont à la fois des modèles à hasard proportionnels et à durée de vie accélérée. La procédure LIFEREG permet de les estimer aisément comme des modèles à durée de vie accélérée. On retrouve les paramètres de l'autre forme moyennant une simple règle de trois.

$$\bar{F}(t) = 1 - \Phi(\frac{\log(t) - xb}{\sigma})$$

On a directement les bons estimateurs

Mise en oeuvre simplifiée (principales options).

PROC LIFEREG < Options 1 > ;
MODEL response = indépendants < 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).
D = ; précise la distribution.
EXPONENTIAL modèle exponentiel
WEIBULL Weibull
LLOGISTIC log Logistique
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 = xb.

. CLASS le même rôle que dans la PROC GLM.

The SAS System

Lifereg Procedure

Data Set =WORK.A
Dependent Variable=Log(Y)
Censoring Variable=C
Censoring Value(s)= 1
Noncensored Values= 1986 Right Censored Values= 828
Left Censored Values= 0 Interval Censored Values= 0

Log Likelihood for WEIBULL -3745.566587] n'est pas égal à l'log_vraisemblance (voir manuel SAS) mais les estimateurs sont les mêmes que pour la "vraie" log.vraisemblance.

Lifereg Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	6.66825994	0.335379	395.325	0.0001	Intercept → $b_0 = -6.99$
AGE	1	2.58784818	0.190872	183.8205	0.0001	→ $b_1 = -2.71$
SCALE	1	0.9551992	0.016359			Extreme value scale parameter

↓ $\alpha = \frac{1}{\sigma} = 1.047$

9.2.2 Utilisation de la PROC NLIN

Un inconvénient déjà cité de la PROC LIFEREG est de pas fournir les estimateurs recherchés, notamment pour le modèle de Weibull, mais des estimateurs divisés par $\hat{\sigma}$. En outre, cette procédure est fermée et ne permet pas de prendre en compte d'autres éléments de la log-vraisemblance, par exemple en présence de sélection endogène. Un moyen d'obtenir les résultats recherchés consiste à utiliser la PROC NLIN, en la paramétrisant afin qu'elle maximise la log-vraisemblance requise. Cette procédure permet en effet de maximiser une fonction quelconque une fois définies cette fonction (appelée fonction de perte, repérée par l'instruction _loss_) et sa dérivée. On fait alors exécuter à la PROC NLIN un algorithme de Gauss Newton (voir supra) dans lequel le Hessien $\frac{\partial^2 LogL}{\partial b \partial b'}$ a été remplacé par son équivalent asymptotique, l'opposé de l'espérance du produit des dérivées premières $-E[\frac{\partial LogL}{\partial b}(\frac{\partial LogL}{\partial b})']$. Dans ce cas, l'instruction MODEL devient inopérante et doit être remplacée afin de générer systématiquement un résidu de 1 afin de générer un score adéquat.

Exemple d'estimation avec une fonction de Hasard Weibull

```
proc nlin data=a sigsq=1 method=marquardt;
Parms a=0.50 b0=0 b1=0;
_xb_=b0+age*b1;
_lsurv_=(y**a)*exp(_xb_);
_loss_=-( (_xb_+log(a)+(a-1)*logy)*d-_lsurv_ );
der.b0= ( d-_lsurv_ );
der.b1=(der.b0)*age;
der.a=  ((1/a + logy)*d-logy*_lsurv_);
model y=y-1;
```

→ nécessaire pour effectuer un E.M.V.

→ vraisemblance

} gradient

↳ instruction facultative destinée à construire une variable de travail

The SAS System 09:54 Tuesday, Oc

Non-Linear Least Squares Iterative Phase Dependent Variable Y Method: Marquardt

Iter	A	B0	B1	Sum of Loss
0	0.500000	0	0	8907.485040
1	0.726234	-0.555944	0.174401	7162.024178
2	1.167913	-2.439682	-0.068306	6400.413907
3	1.103386	-3.641221	-0.814720	6294.941242
4	1.082431	-5.186670	-1.687665	6232.096426
5	1.055668	-6.322047	-2.344031	6210.259140
6	1.052072	-6.862114	-2.640263	6206.997045
7	1.046155	-6.960325	-2.700334	6206.805315
8	1.048412	-6.986344	-2.710340	6206.795405
9	1.045939	-6.977509	-2.708854	6206.793843
10	1.047751	-6.984691	-2.710024	6206.793121
11	1.046240	-6.978401	-2.708780	6206.792662
12	1.047443	-6.983242	-2.709649	6206.792371
13	1.046469	-6.979260	-2.708903	6206.792180
14	1.047252	-6.982442	-2.709489	6206.792057
15	1.046620	-6.979869	-2.709012	6206.791976
16	1.047129	-6.981940	-2.709395	6206.791924

} itérations successives

terion met.

Non-Linear Least Squares Summary Statistics Dependent Variable Y

Source	DF	Sum of Squares	Mean Square
Regression	3	215156.00000	71718.66667
Residual	2811	2814.00000	1.00107
Uncorrected Total	2814	217970.00000	
(Corrected Total)	2813	109487.13433	
Sum of Loss		6206.79192	

} calculs parasites

→ - log vraisemblance

Parameter		Estimate	Asymptotic Std. Error	Asymptotic 95 % Confidence Interval Lower	Upper
α	A	1.047129225	0.02283216945	1.0023589197	1.0918995312
b_0	B0	-6.981939868	0.33730269406	-7.6433375854	-6.3205421512
b_1	B1	-2.709395068	0.18420119775	-3.0705847468	-2.3482053895

Asymptotic Correlation Matrix

Corr	A	B0	B1
A	1	-0.301809437	-0.138226033
B0	-0.301809437	1	0.9840839573
B1	-0.138226033	0.9840839573	1

```
1                                                          The SAS System

NOTE: Copyright(c) 1989 by SAS Institute Inc., Cary, NC USA.
NOTE: SAS (r) Proprietary Software Release 6.07   TS305
      Licensed to INSEE LILLE, FRANCE, Site 0002383002.

NOTE: Running on IBM Model 9121 Serial Number 110337.


  BIENVENUE sous le Systeme SAS release 607
  VERSION EN  PRODUCTION DEPUIS LE 15 MARS 1993


NOTE: The SASUSER library was not specified. SASUSER library will now be the same as the WORK library.
NOTE: All data sets and catalogs in the SASUSER library will be deleted at the end of the session.  Use the NOWORKTERM option to
      prevent their deletion.

NOTE: SAS system options specified are:
      SORT=4 MEMSIZE=30M

NOTE: The initialization phase used 0.16 CPU seconds and 756K.
1          Options Nocaps;
2          data a;set e.th;
3          y=yy;t=ta;
4          logy=log(y);
5          if t>0 then logt=log(t);else logt=0;
6          d=1-cens;
7

NOTE: The data set WORK.A has 3680 observations and 27 variables.
NOTE: The DATA statement used 0.18 CPU seconds and 1099K.

8          proc nlin data=a sigsq=1 method=marquardt;
9          Parms a=0.50 b0=0
10                bcre2=0 bcre3=0 bcre4=0
11                bsant2=0 bsant3=0 bsant4=0 bsant5=0 bsant6=0
12                bage2=0 bage3=0 bage4=0 bage5=0 bage6=0 bage7=0 bage8=0
13                ballocl=0;
14         array ax cre2-cre4 sant2-sant6 age2-age8 allocl;
15         array ab bcre2-bcre4 bsant2-bsant6 bage2-bage8 ballocl;
16         array ad der.bcre2-der.bcre4 der.bsant2-der.bsant6
17                  der.bage2-der.bage8 der.ballocl;
18         _xb_=b0;do over ax;_xb_=_xb_+ax*ab;end;
19         _lsurv_=(y**a)*exp(_xb_);
20         _lsurv0_=(t**a)*exp(_xb_);
21         _loss_=-( (_xb_+log(a)+(a-1)*logy)*d-_lsurv_+_lsurv0_ );
22         der.b0= ( d-_lsurv_+_lsurv0_ );
23         do over ad;
24            ad=(der.b0)*ax;
25         end;
26         der.a=  ((1/a + logy)*d-logy*_lsurv_+logt*_lsurv0_);
27         model y=y-1;
NOTE: PROC NLIN grid search time was  0: 0: 3.
NOTE: PROC NLIN execution time was  0: 3:50.
NOTE: The PROCEDURE NLIN printed pages 1-3.
NOTE: The PROCEDURE NLIN used 19.78 CPU seconds and 1263K.

NOTE: The SAS session used 20.13 CPU seconds and 1263K.
NOTE: SAS Institute S.A., Boite Postale 5, 77166 Gregy-Sur-Yerres
```

Conditionnement par B suivi en

Exemple d'utilisation de NLIN pour estimer

The SAS System

Non-Linear Least Squares Iterative Phase Dependent Variable Y Method: Marquardt

Iter	A BSANT4 BAGE6	B0 BSANT5 BAGE7	BCRE2 BSANT6 BAGE8	BCRE3 BAGE2 BALLOC1	BCRE4 BAGE3	BSANT2 BAGE4	BSANT3 BAGE5	Sum of Loss
0	0.500000	0	0	0	0	0	0	9901.112126
	0	0	0	0	0	0	0	
	0	0	0	0				
1	0.639119	-0.660212	-0.145124	-0.141295	0.237580	-0.023353	0.065083	8927.620053
	0.031295	0.121768	0.053060	-0.117099	-0.098363	-0.201913	-0.178264	
	-0.166989	-0.286190	-0.298007	-0.045502				
2	0.937159	-1.878259	-0.354960	-0.310498	0.598386	0.064612	0.307097	8469.578056
	0.249309	0.487130	0.164600	-0.364173	-0.441418	-0.615229	-0.627447	
	-0.716209	-1.055043	-1.080362	-0.008113				
3	0.916085	-1.840788	-0.325620	-0.256632	0.629418	0.165736	0.400500	8442.037348
	0.406455	0.618585	0.227385	-0.320970	-0.531591	-0.631808	-0.721405	
	-0.872690	-1.395535	-1.505878	0.007059				
4	0.936042	-1.924070	-0.313132	-0.261466	0.638237	0.152028	0.386765	8440.668995
	0.393083	0.612341	0.216761	-0.312841	-0.516456	-0.631182	-0.711143	
	-0.877049	-1.452859	-1.574241	0.008288				
5	0.925767	-1.884176	-0.308677	-0.254251	0.636393	0.158747	0.388150	8440.526502
	0.394547	0.612117	0.218169	-0.315263	-0.517036	-0.631829	-0.715722	
	-0.875525	-1.466164	-1.579630	0.006669				
6	0.930789	-1.903764	-0.309260	-0.256070	0.636459	0.155643	0.387355	8440.495507
	0.393187	0.611893	0.215746	-0.315080	-0.517391	-0.631980	-0.714561	
	-0.877622	-1.467057	-1.580060	0.008299				
7	0.928200	-1.893817	-0.308741	-0.255127	0.636224	0.157277	0.387598	8440.487981
	0.393685	0.611919	0.216676	-0.315084	-0.517199	-0.631690	-0.715007	
	-0.876507	-1.467554	-1.579863	0.007431				
8	0.929471	-1.898685	-0.308930	-0.255488	0.636245	0.156460	0.387472	8440.486132
	0.393381	0.611861	0.216107	-0.315140	-0.517384	-0.631848	-0.714809	
	-0.877113	-1.467499	-1.580002	0.007891				
9	0.928830	-1.896234	-0.308825	-0.255314	0.636209	0.156873	0.387526	8440.485674
	0.393519	0.611880	0.216376	-0.315108	-0.517303	-0.631753	-0.714894	
	-0.876804	-1.467544	-1.579922	0.007655				
10	0.929147	-1.897446	-0.308873	-0.255394	0.636219	0.156666	0.387499	8440.485560
	0.393446	0.611867	0.216235	-0.315128	-0.517352	-0.631800	-0.714852	
	-0.876961	-1.467527	-1.579963	0.007773				
11	0.928989	-1.896840	-0.308848	-0.255355	0.636212	0.156769	0.387512	8440.485532
	0.393481	0.611873	0.216304	-0.315118	-0.517329	-0.631776	-0.714872	
	-0.876883	-1.467536	-1.579941	0.007714				

NOTE: Convergence criterion met.

Itération

Non-Linear Least Squares Summary Statistics Dependant Variable Y

Source	DF	Sum of Squares	Mean Square
Regression	18	2776365.0000	154242.5000
Residual	3662	3680.0000	1.0049
Uncorrected Total	3680	2780045.0000	
(Corrected Total)	3679	1429165.3671	
Sum of Loss		8440.4855	

→ - log. vraisemblance

The SAS System

Parameter	Estimate	Asymptotic Std. Error	Asymptotic 95 % Confidence Interval Lower	Upper
A	0.928988642	0.01679714612	0.8960553636	0.9619219200
B0	-1.896840082	0.08869075323	-2.0707313626	-1.7229488017
BCRE2	-0.308848120	0.06189730114	-0.4302068927	-0.1874893475
BCRE3	-0.255354535	0.07118437333	-0.3949219805	-0.1157870904
BCRE4	0.636211648	0.05026439035	0.5376609097	0.7347623861
BSANT2	0.156769478	0.07291495606	0.0138089709	0.2997299855
BSANT3	0.387511728	0.07190195432	0.2465373595	0.5284860960
BSANT4	0.393481021	0.06433264540	0.2673473974	0.5196146446
BSANT5	0.611872603	0.07321428913	0.4683252090	0.7554199967
BSANT6	0.216304308	0.06774999730	0.0834704630	0.3491381533
BAGE2	-0.315117869	0.05908036215	-0.4309536178	-0.1992821199
BAGE3	-0.517329103	0.06293084475	-0.6407142904	-0.3939439164
BAGE4	-0.631775539	0.07156909241	-0.7720972828	-0.4914537959
BAGE5	-0.714871720	0.08309325626	-0.8777882836	-0.5519551566
BAGE6	-0.876882805	0.08800791202	-1.0494352748	-0.7043303353
BAGE7	-1.467535725	0.09665347325	-1.6570390907	-1.2780323597
BAGE8	-1.579941357	0.10012624833	-1.7762536097	-1.3836291050
BALLOC1	0.007713562	0.05544145467	-0.1009875734	0.1164146967

estimations

Asymptotic Correlation Matrix

Corr	A	B0	BCRE2	BCRE3	BCRE4	BSANT2	BSANT3	BSANT4	BSANT5
A	1	-0.60824987	-0.06172492	-0.04620204	0.081961965	-0.02444655	0.053688745	0.032468422	0.070881502
B0	-0.60824987	1	-0.13171071	-0.23952389	-0.34257101	-0.27788346	-0.30990535	-0.33204981	-0.28984235
BCRE2	-0.06172492	-0.13171071	1	0.371228597	0.546686421	-0.02573659	-0.06946145	0.005611338	-0.07575261
BCRE3	-0.04620204	-0.23952389	0.371228597	1	0.457958334	0.013688122	0.007785048	0.025482736	-0.01012895
BCRE4	0.081961965	-0.34257101	0.546686421	0.457958334	1	-0.00248192	0.005319205	0.019032348	-0.01380476
BSANT2	-0.02444655	-0.27788346	-0.02573659	0.013688122	-0.00248192	1	0.477478414	0.530725712	0.472889249
BSANT3	0.053688745	-0.30990535	-0.06946145	0.007785048	0.005319205	0.477478414	1	0.558720502	0.513463854
BSANT4	0.032468422	-0.33204981	0.005611338	0.025482736	0.019032348	0.530725712	0.558720502	1	0.57109521
BSANT5	0.070881502	-0.28984235	-0.07575261	-0.01012895	-0.01380476	0.472889249	0.513463854	0.57109521	1
BSANT6	-0.08258511	-0.31930292	-0.06406762	-0.14500002	-0.07955073	0.466562859	0.465860891	0.519191492	0.460731824
BAGE2	-0.15298561	-0.07497915	-0.01500272	-0.01040478	-0.01018628	-0.07595395	-0.15328812	-0.12129521	-0.09948079
BAGE3	-0.20072141	-0.00364383	-0.05462895	0.01522364	-0.00594329	-0.0400145	-0.10849829	-0.17550166	-0.16927068
BAGE4	-0.18236084	-0.01453258	-0.12898394	-0.02172015	-0.03009258	0.004498939	-0.09031632	-0.08956218	-0.12694187
BAGE5	-0.15446322	0.008848179	-0.14251663	0.004386751	-0.01235297	-0.02300854	-0.07916651	-0.11655063	-0.18006147
BAGE6	-0.20003946	0.046220298	-0.09752568	0.002693507	-0.0369603	-0.02089758	-0.0892705	-0.10180999	-0.1732832
BAGE7	-0.18145882	0.042149166	-0.12005687	-0.0011079	-0.01424007	0.005239762	-0.05897986	-0.09195843	-0.10850112
BAGE8	-0.17930378	0.048504181	-0.17681905	-0.01426501	-0.01939413	-0.03370265	-0.08067627	-0.07586762	-0.17284731
BALLOC1	0.075773879	-0.4308824	-0.13388365	0.042018526	-0.13189607	-0.12121762	-0.10442226	-0.13713893	-0.11522192

The SAS System

Asymptotic Correlation Matrix

Corr	BSANT6	BAGE2	BAGE3	BAGE4	BAGE5	BAGE6	BAGE7	BAGE8	BALLOC1
A	-0.08258511	-0.15298561	-0.20072141	-0.18236084	-0.15446322	-0.20003946	-0.18145882	-0.17930378	0.075773879
B0	-0.31930292	-0.07497915	-0.00364383	-0.01453258	0.008848179	0.046220298	0.042149166	0.048504181	-0.4308824
BCRE2	-0.06406762	-0.01500272	-0.05462895	-0.12898394	-0.14251663	-0.09752568	-0.12005687	-0.17681905	-0.13388365
BCRE3	-0.14500002	-0.01040478	0.01522364	-0.02172015	0.004386751	0.002693507	-0.0011079	-0.01426501	0.042018526
BCRE4	-0.07955073	-0.01018628	-0.00594329	-0.03009258	-0.01235297	-0.0369603	-0.01424007	-0.01939413	-0.13189607
BSANT2	0.466562859	-0.07595395	-0.0400145	0.004498939	-0.02300854	-0.02089758	0.005239762	-0.03370265	-0.12121762
BSANT3	0.465860891	-0.15328812	-0.10849829	-0.09031632	-0.07916651	-0.0892705	-0.05847986	-0.08067627	-0.10442226
BSANT4	0.519191492	-0.12129521	-0.17550166	-0.08956218	-0.11655063	-0.10180999	-0.09195843	-0.07586762	-0.13713893
BSANT5	0.460731824	-0.09948079	-0.16927068	-0.12694187	-0.18006147	-0.1732832	-0.10850112	-0.17284731	-0.11522192
BSANT6	1	0.035438379	0.024103824	0.044338329	0.017150947	0.022057617	0.020043808	-0.00057826	0.030491443
BAGE2	0.035438379	1	0.368902838	0.322045832	0.284663236	0.275558207	0.248395244	0.244098422	-0.01495185
BAGE3	0.024103824	0.368902838	1	0.341194016	0.317213429	0.303325407	0.276470095	0.274077547	-0.07897482
BAGE4	0.044338329	0.322045832	0.341194016	1	0.287122066	0.272226878	0.250148447	0.257877598	-0.04561988
BAGE5	0.017150947	0.284663236	0.317213429	0.287122066	1	0.258067704	0.236551464	0.249737012	-0.06519102
BAGE6	0.022057617	0.275558207	0.303325407	0.272226878	0.258067704	1	0.223501445	0.23314897	-0.06707533
BAGE7	0.020043808	0.248395244	0.276470095	0.250148447	0.236551464	0.223501445	1	0.214516795	-0.08481245
BAGE8	-0.00057826	0.244098422	0.274077547	0.257877598	0.249737012	0.23314897	0.214516795	1	-0.05317916
BALLOC1	0.030491443	-0.01495185	-0.07897482	-0.04561988	-0.06519102	-0.06707533	-0.08481245	-0.05317916	1

9.3 Estimations semi-paramétriques, la procédure PHREG

Elle 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 covariable affectant le hasard de base sous la forme $\exp(x\beta)$.

Mise en oeuvre simplifiée:

PROC PHREG	< Options 1 >;
MODEL	*time* * *flag*( ) = *exogènes*;
FREQ	*variable*(*entière*);
OUTPUT	< Options 2 >;
BASELINE	< Options 3 >;

Options 1:

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

Options 2:

OUT= *data*	nom du *data* de sortie construit à partir du tableau initial et contenant les statistiques requises .
XBETA =	$x\beta$
SURVIVAL	survie
LOGSURV	Log(survie)

Options 3:

OUT= *data*	nom du *data* de sortie contenant la valeur de la survie pour les valeurs possibles de *time*
COVARIATES= *data*	nom du *data* contenant les valeurs des covariables pour lesquelles on cherche à calculer la survie (par défaut, SAS prend les valeurs moyennes de ces covariables dans le fichier)
XBETA=	$x\beta$
SURVIVAL	survie
LOGSURV	Log(survie)

The SAS System

The PHREG Procedure

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

```
14          proc phreg data=a;
15          model dur*c(1)=femmes;
16          baseline  out=b covariates=cov survival=s logsurv=ls;
NOTE: The PROCEDURE PHREG used 4.85 CPU seconds and 3171K.
NOTE: The data set WORK.B has 105 observations and 4 variables.
NOTE: The PROCEDURE PHREG printed page 1.
```

2 observations dans cov

dos	Femmes	
1	0	→ hommes
2	1	→ femmes

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
25046	13698	11348	45.31

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	257467.329	257396.229	71.100 with 1 DF (p=0.0001)
Score	.	.	71.435 with 1 DF (p=0.0001)
Wald	.	.	71.311 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
FEMMES	1	-0.144535	0.01712	71.31139	0.0001	0.865

56

Sous la forme de exp(-0,144) dans la translation sur la hazard de base entre femmes et femmes.

The SAS System

OBS	FEMMES	DUR	S	LS
1	0	0	1.00000	0.00000
2	0	0	0.86263	-0.14777
3	0	1	0.74747	-0.29105
4	0	2	0.66471	-0.40841
5	0	3	0.59660	-0.51651
6	0	4	0.53475	-0.62596
7	0	5	0.48380	-0.72608
8	0	6	0.44450	-0.81081
9	0	7	0.40534	-0.90304
10	0	8	0.37649	-0.97686
11	0	9	0.34752	-1.05692
12	0	10	0.30782	-1.17825
13	0	11	0.28266	-1.26351
14	0	12	0.26013	-1.34657
15	0	13	0.24304	-1.41455
16	0	14	0.22649	-1.48504
17	0	15	0.20933	-1.56385
18	0	16	0.19677	-1.62572
19	0	17	0.18553	-1.68456
20	0	18	0.17605	-1.73701
21	0	19	0.16267	-1.81603
22	0	20	0.15181	-1.88510
23	0	21	0.14012	-1.96527
24	0	22	0.13316	-2.01623
25	0	23	0.12786	-2.05681
26	0	24	0.12105	-2.11157
27	0	25	0.11787	-2.13821
28	0	26	0.10999	-2.20734
29	0	27	0.10280	-2.27497
30	0	28	0.09838	-2.31891
31	0	29	0.09496	-2.35430
32	0	30	0.08983	-2.40986
33	0	32	0.08519	-2.46292
34	0	33	0.08199	-2.50117
35	0	34	0.07561	-2.58213
36	1	0	1.00000	0.00000
37	1	0	0.87996	-0.12788
38	1	1	0.77733	-0.25189
39	1	2	0.70226	-0.35344
40	1	3	0.63955	-0.44700
41	1	4	0.58175	-0.54172
42	1	5	0.53346	-0.62837
43	1	6	0.49575	-0.70169
44	1	7	0.45771	-0.78151
45	1	8	0.42939	-0.84539
46	1	9	0.40064	-0.91469
47	1	10	0.36071	-1.01969
48	1	11	0.33505	-1.09347
49	1	12	0.31181	-1.16536
50	1	13	0.29400	-1.22418
51	1	14	0.27660	-1.28519
52	1	15	0.25836	-1.35340
53	1	16	0.24489	-1.40693
54	1	17	0.23273	-1.45786
55	1	18	0.22241	-1.50325
56	1	19	0.20770	-1.57164

Calcul de $\hat{S}_0$ associé et de son log pour 3 sous-populations

femmes ($X=0$)

femmes ($X=1$)

The SAS System

OBS	FEMMES	DUR	S	LS
57	1.00000	20	0.19565	-1.63141
58	1.00000	21	0.18254	-1.70079
59	1.00000	22	0.17466	-1.74490
60	1.00000	23	0.16864	-1.78001
61	1.00000	24	0.16083	-1.82741
62	1.00000	25	0.15716	-1.85046
63	1.00000	26	0.14804	-1.91029
64	1.00000	27	0.13962	-1.96881
65	1.00000	28	0.13441	-2.00684
66	1.00000	29	0.13036	-2.03747
67	1.00000	30	0.12424	-2.08555
68	1.00000	32	0.11866	-2.13147
69	1.00000	33	0.11480	-2.16457
70	1.00000	34	0.10703	-2.23464
71	0.53478	0	1.00000	0.00000
72	0.53478	0	0.87217	-0.13678
73	0.53478	1	0.76383	-0.26941
74	0.53478	2	0.68521	-0.37803
75	0.53478	3	0.61997	-0.47809
76	0.53478	4	0.56023	-0.57940
77	0.53478	5	0.51065	-0.67207
78	0.53478	6	0.47213	-0.75050
79	0.53478	7	0.43350	-0.83587
80	0.53478	8	0.40487	-0.90419
81	0.53478	9	0.37595	-0.97831
82	0.53478	10	0.33601	-1.09061
83	0.53478	11	0.31052	-1.16952
84	0.53478	12	0.28753	-1.24641
85	0.53478	13	0.27000	-1.30933
86	0.53478	14	0.25295	-1.37458
87	0.53478	15	0.23515	-1.44753
88	0.53478	16	0.22206	-1.50479
89	0.53478	17	0.21029	-1.55926
90	0.53478	18	0.20033	-1.60780
91	0.53478	19	0.18620	-1.68095
92	0.53478	20	0.17467	-1.74488
93	0.53478	21	0.16217	-1.81909
94	0.53478	22	0.15470	-1.86626
95	0.53478	23	0.14900	-1.90382
96	0.53478	24	0.14163	-1.95451
97	0.53478	25	0.13818	-1.97917
98	0.53478	26	0.12962	-2.04316
99	0.53478	27	0.12175	-2.10575
100	0.53478	28	0.11690	-2.14642
101	0.53478	29	0.11313	-2.17918
102	0.53478	30	0.10746	-2.23061
103	0.53478	32	0.10231	-2.27972
104	0.53478	33	0.09875	-2.31513
105	0.53478	34	0.09162	-2.39007

$\bar{x} = 0.53$: moyenne de x dans Richardson (53% de femmes)

10 Pour en savoir plus

Des éléments théoriques plus détaillés sur les modèles de durée figurent dans les documents ci-dessous. Vous y trouverez en particulier les démonstrations des propriétés énoncées dans ce fascicule, des propositions de tests supplémentaires, et d'autres exemples d'applications,...ainsi qu'une bibliographie plus complète.

- **J.J. DROESBEKE, B. FICHET, P. TASSI**, "Analyse statistique des durées de vie-Modélisation des données censurées ", *Economica*, 1989.

- **C GOURIEROUX**," Econométrie des variables qualitatives", *Economica*, 1989.

- **T. LANCASTER**, "The Econometric Analysis of Transition Data", *Econometric Society Monographs*, Cambridge University Press, 1990.

- **A. MOREAU**, "Econométrie des variables de durée", **Note Département recherche** N.123/G 305, 1989.

Les descriptions complètes des procédures SAS présentées figurent bien entendu, avec des exemples supplémentaires, dans les manuels de référence de SAS-V6. La procédure PHREG fait l'objet d' un fascicule spécifique.