MODÈLES DE DURÉE DE VIE

I. Généralités

On rappelle tout d'abord quelques fonctions associées à une durée de vie et leurs propriétés.

La loi de distribution d'une variable aléatoire (v.a.) continue à densité est totalement caractérisée par celle-ci, ou tout aussi bien par la fonction de répartition associée.

Dans le cas d'une v.a. modélisant une ***durée de vie*** (supposée continue à densité), on a coutume, en démographie mathématique, de considérer d'autres fonctions également caractéristiques, et pour certaines d'interprétation plus immédiate.

Soit T une v.a. de durée de vie. On désignera par:

- $f(t)$ la densité, à valeurs dans $\mathbf{R}^+$ ($\int_0^\infty f(x).dx = 1$).
- $F(t) = P(T < t) = \int_0^t f(x).dx$ la ***fonction de répartition***, qui mesure la probabilité de "mourir" au plus tard en t.
- $S(t) = P(T \geq t) = 1 - F(t) = \int_t^\infty f(x).dx$ la ***fonction de survie***, qui mesure la probabilité de survivre à t.
- $h(t) = f(t)/S(t) = -d[\ln(S)]/dt$ la ***fonction de risque,*** *hazard function* pour les anglo-saxons et parfois *force de mortalité* (désuet). Elle s'interprète comme la densité de mortalité en t conditionnée par la survie jusque-là.
- $H(t) = \int_0^t h(x).dx$ la ***fonction de risque cumulé***.
- $e(t)$ l'***espérance de survie*** si on a vécu jusqu'à t.

N'importe laquelle de ces fonctions est caractéristique de la distribution. Voici quelques unes des autres relations qui les lient.

- $S(t) = \exp[-\int_0^t h(x).dx]$ ou $h(t) = d[-\ln(S(t)]/dt$
- $H(t) = -\ln[S(t)]$
- $e(t) = [\int_t^\infty S(x).dx]/ S(t)$

II. Distributions classiques

On présente à titre d'exemples quelques familles de distributions usuelles avec leurs propriétés.

Distribution exponentielle (un paramètre λ, positif)

$f(t) = \lambda \exp(-\lambda.t)$
$S(t) = \exp(-\lambda.t)$

$h(t) = \lambda$
$H(t) = \lambda.t$

C'est la distribution à risque constant ou *sans mémoire*, cette propriété est caractéristique; il est équivalent de dire que le logarithme de la fonction de survie: LS = ln(S), est linéaire.

Il est par ailleurs à noter que le mélange de distributions exponentielles ne donne pas une distribution exponentielle (de même que celui de lois normales ne donne pas une loi normale).

Distribution de Weibull (deux paramètres α et ß, positifs)

$f(t) = \alpha.ß.(ß.t)^{\alpha-1}.\exp[-(ß.t)^{\alpha}]$
$S(t) = \exp[-(ß.t)^{\alpha}]$
$h(t) = \alpha.ß.(ß.t)^{\alpha-1}$
$H(t) = (ß.t)^{\alpha}$

Pour $\alpha = 1$ on retrouve la distribution exponentielle.

Comme pour celle-ci, la fonction de risque est monotone.

Le logarithme de l'antilogarithme de la fonction de survie: ln[-ln(S)] parfois abusivement noté LLS, est une fonction affine; cette propriété est caractéristique.

Distribution log-normale (deux paramètres α et ß, positifs)

C'est la loi d'une v.a. dont le logarithme suit une loi $N(\ln(1/\alpha),ß)$

$$f(t) = \frac{1}{ß.t.(2\Pi)^{1/2}}.\exp\{-\frac{[\ln(t.\alpha)]^2}{2.ß^2}\}$$

$S(t) = 1 - \Phi[\ln(t.\alpha)/ß]$ où Φ est la fonction de répartition de la loi N(0,1)

Les démographes utilisent parfois des distributions paramétriques plus particulières (telle celle de Coale-Trussel) permettant de mieux approcher les singularités de l'existence humaine.

III. Estimation des modèles

Les modèles de durée se prêtent à divers types d'estimation:

- L'estimation ***fonctionnelle*** (ou ***non paramétrique***), qui vise à approximer l'une ou plusieurs des différentes fonctions caractérisant la distribution observée (F ou h le plus souvent) sans faire d'hypothèse sur celle-ci.

- L'estimation ***paramétrique***, qui ayant retenu une forme de distribution donnée (par exemple la loi exponentielle ou la loi de Weibull) cherche à en estimer les paramètres. Un terme correctif pouvant prendre en compte l'effet de variables exogènes ou *covariables*.

Ex.1 ***modèle à temps accéléré***: $S(t/x) = S_0(t.e^{ß.x})$ où S_0 représente la fonction de survie de base retenue, x un vecteur de covariables et ß les coefficients associés.

Ex.2 ***modèle à risque proportionnel***: $h(t/x) = h_0(t).e^{ß.x}$ où h_0 est la fonction de risque de base retenue.

- L'estimation ***semi-paramétrique***, qui pour des modèles de la forme précédente cherche à estimer l'influence des facteurs exogènes sans hypothèse concernant la distribution de base.

Une difficulté supplémentaire provient du fait que les données peuvent être ***tronquées*** ou ***censurées***. Pour nous limiter à ce cas, on dira qu'une observation de durée de vie est *censurée à droite* si on connaît non la date de mort, mais simplement une date de dernière observation du sujet vivant; ce serait renoncer à une part d'information que d'écarter une telle observation, son exploitation demande néanmoins un traitement particulier.

Le modèle à temps accéléré

Ce modèle suppose que la fonction de survie S(t) conditionnée par les variables exogènes que nous désignons globalement par x, se ramène à une fonction de survie de base $S_0(t)$, selon une relation:

$$S(t/x) = S_0(t.e^{ß.x})$$

où ß désigne le vecteur des coefficients associés aux variables.

L'estimation d'un tel modèle demande que soit spécifiée la distribution de base, elle opère par la méthode classique du maximum de vraisemblance.

Le modèle à risque proportionnel

Ce modèle, introduit par ***Cox***, suppose que la fonction de risque h(t) conditionnée par les variables exogènes x se ramène à une fonction de risque de base $h_0(t)$ selon une relation:

$$h(t/x) = h_0(t).e^{ß.x}$$

c'est le risque lui-même qui est "modulé" en fonction des exogènes.

Ce modèle qui en général n'est pas équivalent au précedent présente les particularités suivantes:

- Il est caractérisé par des courbes LLS (ou ln[-ln(S)], logarithme de l'antilogarithme de la fonction de survie) parallèles pour les diverses combinaisons de valeurs des covariables, ce qui permet une identification géométrique sur ces courbes estimées.

- Il est possible d'estimer les coefficients ß sans faire d'hypothèse sur la forme de h_0 (il est néanmoins possible d'estimer la distribution de base).

- On peut inclure dans les exogènes des variables dépendant du temps, dont la significativité éventuelle permet alors de récuser le modèle.

Le cas le plus simple est celui ou les exogènes se réduisent à une variable indicatrice, permettant ainsi de tester l'homogénéité de deux sous-populations.

PROCÉDURES SAS

Procédure LIFEREG

La procédure **LIFEREG** permet l'ajustement paramétrique des modèles de durée de vie, c'est à dire l'ajustement à des lois standard classiques. Elle accepte les observations censurées et autorise l'emploi de covariables. Ainsi, pour le modèle classique à temps accéléré, une forme fonctionnelle de fonction de survie S étant retenue, la probabilité de survie à t en présence de covariables notées globalement x est donnée par :

$$S(t/x) = S_0(t.e^{ß.x})$$

et la procédure estime les paramètres de la distribution retenue ainsi que les paramètres ß associés aux covariables.

La syntaxe de base est :

PROC LIFEREG;
MODEL durée [***** censure(liste)] = covariables [/ options];
[**CLASS** variables;]

où *durée* est la durée de vie observée, *censure* une variable indiquant une censure à droite éventuelle notée par les valeurs ou items donnés dans *liste*. Parmi les options, **DIST**=*type* permet de préciser la loi retenue, la loi de **WEIBULL** est prise par défaut, les *types* **EXPONENTIAL**, **GAMMA**, **LOGISTIC**, **LNORMAL** ou **NORMAL** sont également possibles. **CLASS** permet d'indiquer des variables à items qui seront converties en indicatrices et utilisées comme covariables.

Exemple

PROC LIFEREG;
MODEL t ***** sorti(1) = age z / **DIST**=**EXPONENTIAL**;

estime un modèle de survie après une certaine opération chirurgicale, *sorti* étant une variable logique signalant la sortie de l'échantillon et z une autre variable indicatrice indiquant l'administration d'un traitement particulier.

Procédure LIFETEST

La procédure **LIFETEST** permet l'estimation non-paramétrique de fonctions de survie, c'est à dire le calcul des probabilités empiriques de survie. Elle accepte les observations censurées, et peut effectuer des tests d'homogénéité entre groupes et des tests de rang de liaison avec des covariables.

Sa syntaxe de base est :

PROC LIFETEST [options];

TIME durée [***** censure (liste)];
[**FREQ** variables;]
[**STRATA** variables;]
[**TEST** variables;]

parmi les options **PLOTS** = (**S**, **LS**, **LLS**, **H** ou **P**) permet de demander le graphique de la fonction de survie estimée, de son logarithme, de son log-log, de la fonction de risque ("hazard function") ou de la densité. Ces sorties sont fort utiles pour identifier des distributions classiques (qui se prêteront alors à une estimation paramétrique) :

LS linéaire caractérise une loi exponentielle
LLS affine caractérise une loi de Weibull
LLS parallèles signalent un risque proportionnel (dans le cas de deux ou plusieurs groupes)

FREQ est à utiliser en cas de données groupées pour indiquer l'effectif associé à chaque valeur de *durée*. **STRATA** définit le découpage selon lequel seront effectués des calculs séparés puis des tests d'homogénéité. **TEST** enfin indique des covariables quantitatives qui seront comparées à *durée* en des tests de rang.

Exemple

PROC LIFETEST PLOTS = (**S**, **H**);
TIME t * sorti (1);
STRATA sexe;
TEST age delai;

étudie la survie t à une certaine intervention chirurgicale pour chaque sexe, affiche les courbes de survie et de risque, et recherche une liaison avec les covariables *age* et *delai*.

Procédure PHREG

La procédure **PHREG** permet l'estimation du modèle à risque proportionnel :

$$h(t/x) = h_0(t.e^{\beta.x})$$

Sa syntaxe est :

PROC PHREG;
MODEL durée ***** censure (liste) = covariables / options;
[**STRATA** variables];

Exemple

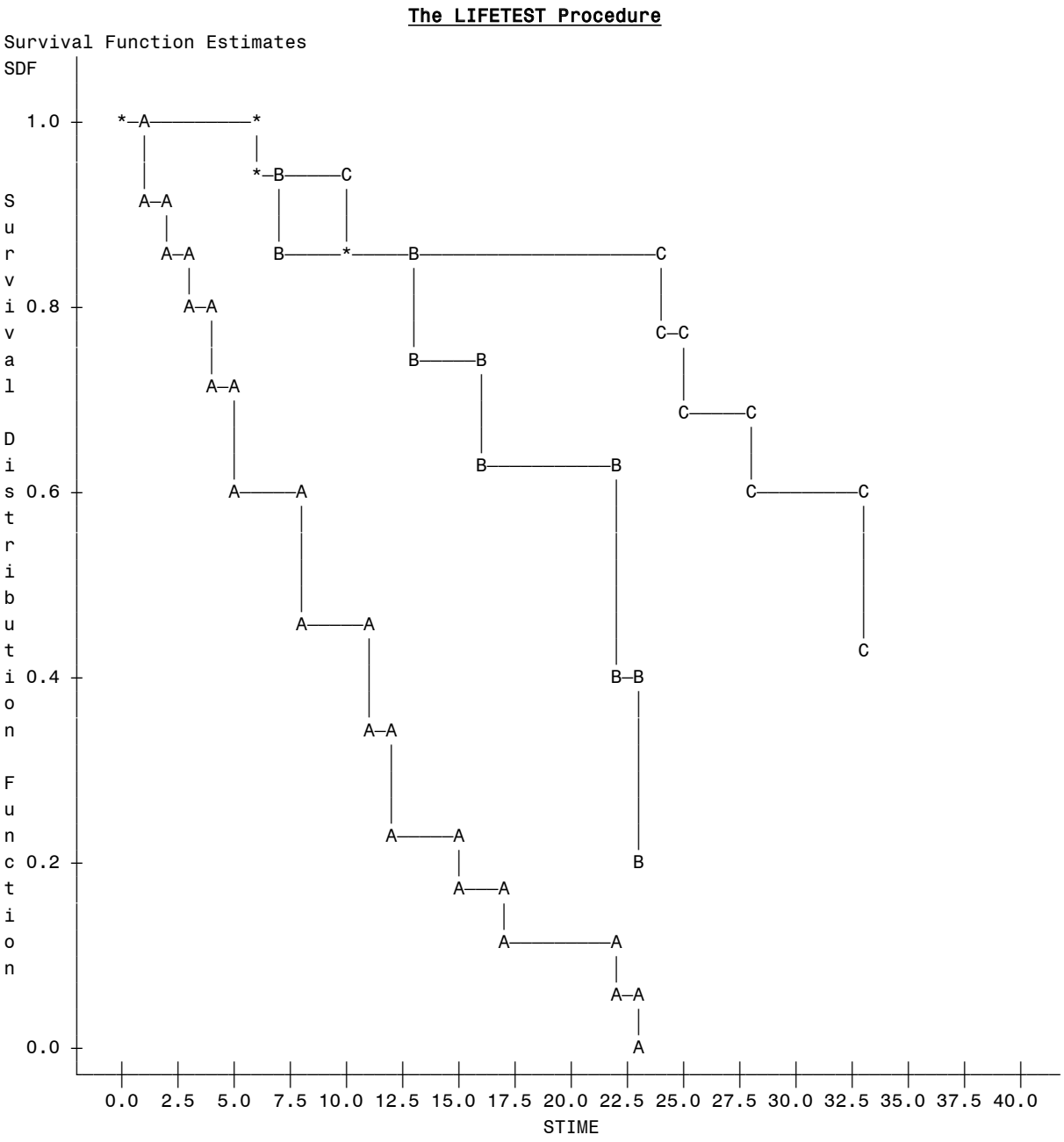
PROC PHREG;
MODEL z ***** parti (1) = x;

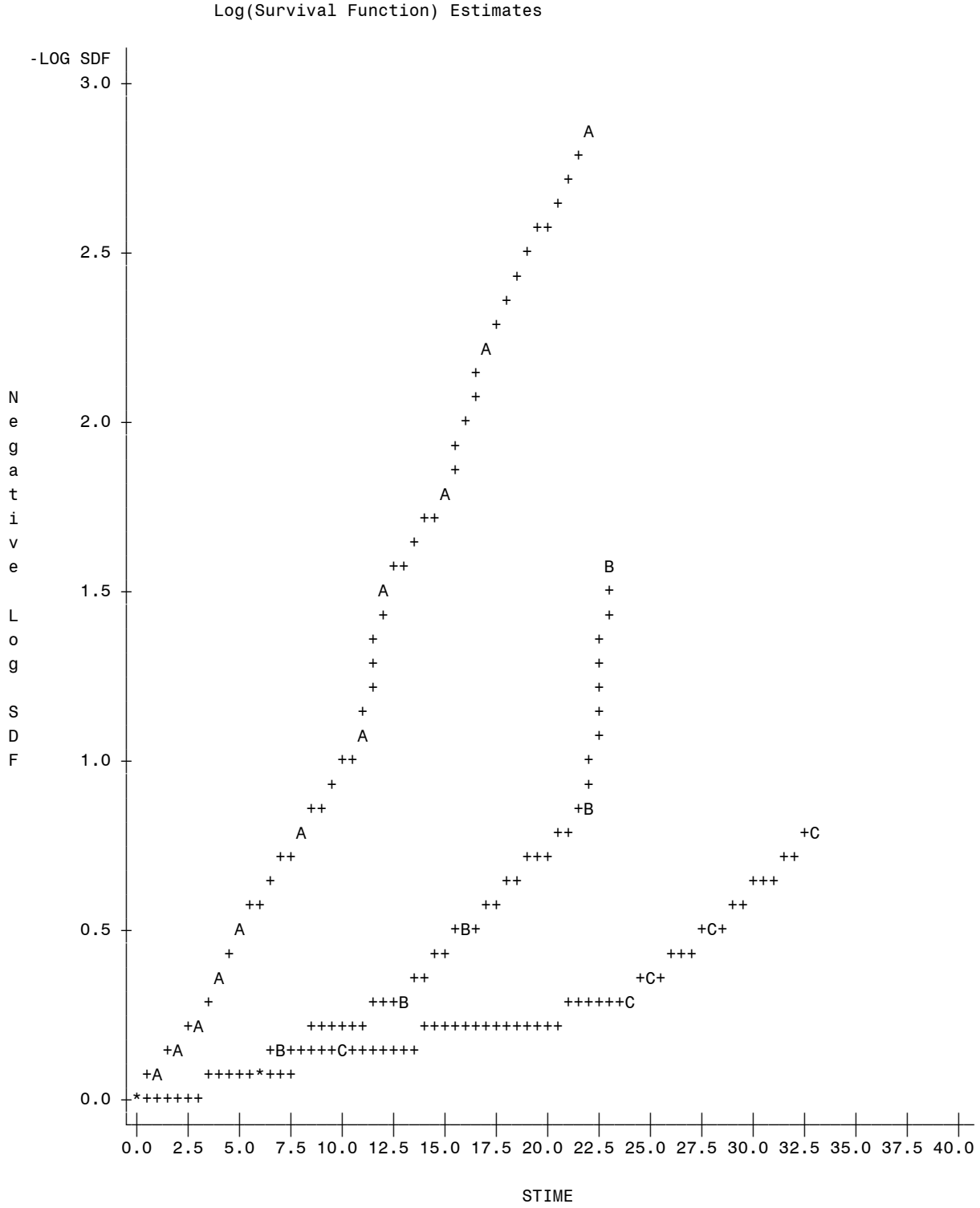
mesure l'effet de la variable indicatrice x indiquant un traitement particulier sur la durée de rémission z d'une certaine affection.

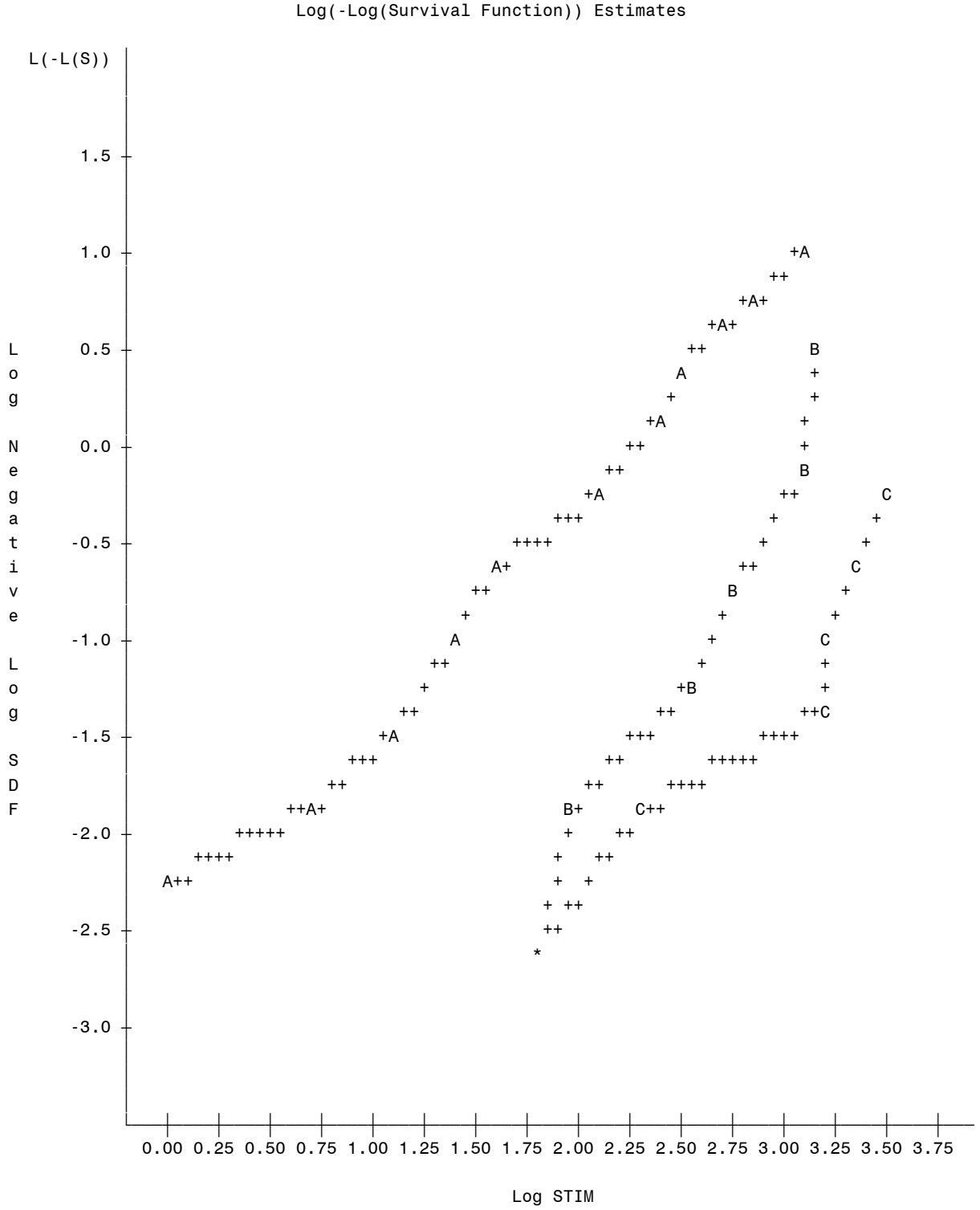
Un exemple, traité à l'aide du logiciel SAS, des méthodes précédentes sur des données médicales (les sorties sont abrégées). La procédure LIFETEST procède d'abord à des estimations non paramétriques, la procédure LIFEREG estime des modèles paramétriques à temps accéléré, et la procédure PHREG estime un modèle à risque proportionnel.

STIME est la variable de durée, DIED note par un 0 les données censurées, DRUG note par un 1 l'administration d'un placebo et par 2 et 3 deux traitements véritables, AGE mesure enfin l'âge des sujets.

```
/* Exemple modèles de durée : LIFETEST, LIFEREG et PHREG */

data cancer;
input stime died drug age;
/* stime = durée de survie              */
/* died : 1 = mort, 0 = censurée        */
/* drug : 1 = placebo, 2 = traitement A */
/*        3 = traitement B              */
cards;
       1         1         1        61
       1         1         1        65
       .         .         .         .
      35         0         3        48
      39         0         3        52
;

proc lifetest plots=(S,LS,LLS);
time stime*died(0);
strata drug;

proc lifereg;
model stime*died(0)=drug age/dist=exponential;
model stime*died(0)=drug age;

proc phreg;
model stime*died(0)=drug age;

run;
```


The LIFETEST Procedure
Survival Function Estimates
SDF
Survival Distribution Function
1.0
0.8
0.6
0.4
0.2
0.0
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0 22.5 25.0 27.5 30.0 32.5 35.0 37.5 40.0
STIME


Log(Survival Function) Estimates
-LOG SDF
Negative Log SDF
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0 22.5 25.0 27.5 30.0 32.5 35.0 37.5 40.0
STIME


Log(-Log(Survival Function)) Estimates
L(-L(S))
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
Log Negative Log SDF
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.25 2.50 2.75 3.00 3.25 3.50 3.75
Log STIM

Testing Homogeneity of Survival Curves over Strata
Time Variable STIME

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	30.1864	2	0.0001
Wilcoxon	23.4533	2	0.0001
-2Log(LR)	20.0245	2	0.0001

The Lifereg Procedure

Data Set =WORK.CANCER
Dependent Variable=Log(STIME)
Censoring Variable=DIED
Censoring Value(s)= 0
Noncensored Values= 31 Right Censored Values= 17
Left Censored Values= 0 Interval Censored Values= 0

Log Likelihood for EXPONENT -48.83759796

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	5.62905462	1.846104	9.297336	0.0023	Intercept
DRUG	1	1.01459184	0.243573	17.35093	0.0001	
AGE	1	-0.078479	0.032154	5.957026	0.0147	
SCALE	0	1	0			Extreme value scale parameter

Lagrange Multiplier ChiSquare for Scale 208.2797 Pr>Chi is 0.0001.

Log Likelihood for WEIBULL -42.66283814

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	5.10129913	1.088864	21.94897	0.0001	Intercept
DRUG	1	0.76994457	0.147806	27.13518	0.0001	
AGE	1	-0.0630939	0.019449	10.52439	0.0012	
SCALE	1	0.56890819	0.079911			Extreme value scale parameter

The PHREG Procedure

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	199.823	163.530	36.293 with 2 DF (p=0.0001)
Score	.	.	33.708 with 2 DF (p=0.0001)
Wald	.	.	27.247 with 2 DF (p=0.0001)

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Risk Ratio
DRUG	1	-1.535422	0.31431	23.86444	0.0001	0.215
AGE	1	0.110065	0.03613	9.27848	0.0023	1.116

----==oΩo==----

(maj 24.02.2009)