Modèle de prévision de défaillance à risques proportionnels : le modèle mODDEF

Olivier Piller

Cemagref
LA RECHERCHE POUR L'INGENIERIE DE L'AGRICULTURE ET DE L'ENVIRONNEMENT

Département Equipements
pour l'Eau et l'Environnement

Division Ouvrages et Réseaux
Hydrauliques

MODELE DE PREVISION DE DEFAILLANCE A RISQUES PROPORTIONNELS : LE MODELE MODDEF

Problématique, Algorithme de résolution et Organigramme

Olivier PILLER

Groupement de Bordeaux

50, Avenue de Verdun
33612 CESTAS Cedex
Tél. 05 5789 08 00 - Fax 05 57 89 08 01

Février 1998

MODELE DE PREVISION DE DEFAILLANCE A RISQUES PROPORTIONNELS

SOMMAIRE

1) Principales notations

T_i	:	V.A. durée de vie jusqu'à la première panne ; t_i une réalisation ;
C_i	:	V.A. durée de l'étude ; c_i une réalisation ;
U_i	:	V.A. $\min(C_i, T_i)$; u_i une réalisation ;
Δ_i	:	V.A. $\begin{cases}1 \text{ mort} \\ 0 \text{ si censure}\end{cases}$; δ_i une réalisation ;
m	:	nombre d'individus ;
s	:	nombre de paramètres de régression ;
$n = s + 2$	:	nombre de paramètres à estimer.

2) Principales définitions

Fonction de survie :

$$S(t) = \text{prob}(T \geq t) = 1 - F(t) \quad 0 < t < \infty$$

$$S(0) = 1 \quad \lim_{t \to \infty} S(t) = 0$$

$$S(t) = \int_t^{\infty} f(x)dx \text{ où f est la fct densité. Donc } S'(t) = -f(t)$$

Fonction risque :

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -(\ln S(t))', \qquad h(t) > 0 \text{ et } \int_0^{\infty} h(u)du = +\infty$$

$$S(t) = e^{-\int_0^t h(u)du}$$

3) Modèle de Weibull à risque proportionnel

« à risque proportionnel » : les covariables agissent proportionnellement sur la fonction risque

$$h(t / \lambda, p, \beta, z) = \lambda p(\lambda t)^{p-1} e^{t z.\beta}$$

$z \in \mathbb{R}^s$: valeur du vecteur des variables explicatives

$\beta \in \mathbb{R}^s$: vecteur des paramètres de la régression

λ : paramètre de position ($\lambda > 0$)

p : paramètre d'échelle ($\sigma > 0$)

Il suit que :

$$S(t) = e^{-(\lambda t)^p e^{\,t z.\beta}}$$

d'où $$f(t) = \lambda p(\lambda t)^{p-1} e^{\,t z.\beta} . e^{-(\lambda t)^p e^{\,t z.\beta}} \qquad \textbf{(1)}$$

Un bon moyen pour vérifier de façon empirique la distribution de Weibull est de tracer : $\ln\left(-\ln(\hat{S}_0(t))\right)$ en fonction de $\ln t$; $\hat{S}_0(t)$ est par exemple l'estimation sur l'échantillon de Kaplan-Meier de la fonction de survie de base (sans les termes de régression).

4) Le modèle log-Linéaire :

On envisage le changement de variable :

$W = \dfrac{\ln(U) - {}^t x.\theta}{\sigma}$, où la loi de U est celle d'une régression de type Weibull (1)

$x \in \mathbb{R}^{s+1}$: $x = \begin{pmatrix} 1 \\ z \end{pmatrix}$

$\theta \in \mathbb{R}^{s+1}$: $\theta = \begin{pmatrix} -\ln(\lambda) \\ -\dfrac{1}{p}\beta \end{pmatrix}$

$\sigma > 0$: $\sigma = p^{-1}$

Calculons la fonction de survie associée à W :

$$S^*(w) = \text{prob}(W \geq w) = \text{prob}\left(\frac{\ln(U) - {}^t x.\theta}{\sigma} \geq w \right) = \text{prob}\left(\ln(U) \geq \sigma w + {}^t x.\theta\right)$$

$$S^*(w) = \text{prob}\left(U \geq e^{\sigma w + {}^t x.\theta} \right) = S(e^{\sigma w + {}^t x.\theta} = u)$$

$$S^*(w) = S(e^{\sigma w + {}^t x.\theta}) = e^{-\left(\lambda e^{p^{-1}.w - \ln(\lambda) - p^{-1}.{}^t z.\beta}\right)^p .e^{{}^t z.\beta}} = e^{-\lambda^p e^{w - \ln(\lambda^p) - {}^t z.\beta} .e^{{}^t z.\beta}} = e^{-e^w}$$

On reconnaît ici la fonction de survie de la loi de Gumbel (dite aussi loi de Gompertz ou double exponentielle ou extreme value distribution), de densité : $e^{w - e^w}$.

donc :

$$E(w) = -e$$
$$var(w) = \frac{\pi^2}{6}$$

5) Construction de la vraisemblance :

Hypothèses :

les C_i sont des V.A. de fonction de survie G(.) et de densité g(.)

$C_1, C_2, ..., C_m, T_1, T_2, ..., T_m$ sont stochastiquement indépendantes

Alors :

$$prob\left(U_i \in [t_i, t_i + dt], \delta_i = 1 / \sigma, \theta, x_i\right) = prob\left(T_i \in [t_i, t_i + dt], C_i > t_i / \sigma, \theta, x_i\right)$$
$$= G(t_i + 0).f(t_i / \sigma, \theta, x_i)dt$$

et

$$prob\left(U_i \in [t_i, t_i + dt], \delta_i = 0 / \sigma, \theta, x_i\right) = prob\left(C_i \in [t_i, t_i + dt], T_i > t_i / \sigma, \theta, x_i\right)$$
$$= g(t_i).\overline{S}(t_i + 0 / \sigma, \theta, x_i)dt$$

S continue, les (U_i, Δ_i) indépendants, et les lois de C_i ne dépendants pas des paramètres :

$$\overline{L}(\sigma, \theta) = \prod_{i=1}^{m} \left[G(t_i + 0).f(t_i / \sigma, \theta, x_i)\right]^{\delta_i} .\left[g(t_i).S(t_i + 0 / \sigma, \theta, x_i)\right]^{1-\delta_i}$$

d'où :

$$\ln \overline{L}(\sigma, \theta) = \sum_{i=1}^{m} \delta_i . \ln f(t_i / \sigma, \theta, x_i) + \sum_{i=1}^{m} (1 - \delta_i).\ln S(t_i / \sigma, \theta, x_i) + Cst \qquad \textbf{(2)}$$

6) Estimations dans un modèle de régression log-linéaire :

on cherche à exprimer la fonction de vraissemblance à partir de $y_i = \ln u_i$

Il suffit de calculer la fonction vraisemblance correspondante :

Sur la V.A. $y_i = \ln t_i$ avec $w_i = \dfrac{y_i - {}^t x_i.\theta}{\sigma}$

La fonction de survie est : $S(y_i) = e^{-e^{w_i}} = S^*(w_i) = S(t_i)$

La densité est $f(y_i) = \dfrac{1}{\sigma} e^{w_i - e^{w_i}} = \dfrac{1}{\sigma} f^*(w_i) = -\left(S(e^{y_i})\right)' = t_i.f(t_i)$

Ainsi à une constante additive près :

$$\ln L(\sigma,\theta) = \sum_{i=1}^{m} \delta_i . \ln f(y_i / \sigma, \theta, x_i) + \sum_{i=1}^{m} (1-\delta_i). \ln S(y_i / \sigma, \theta, x_i)$$

$$\ln L(\sigma,\theta) = -\sum_{i=1}^{m} \delta_i . \ln \sigma + \sum_{i=1}^{m} \delta_i . (w_i - e^{w_i}) + \sum_{i=1}^{m} (1-\delta_i)(-e^{w_i})$$

$$\ln L(\sigma,\theta) = -\left(\sum_{i=1}^{m} \delta_i \right) \ln \sigma + \sum_{i=1}^{m} \delta_i . w_i - \sum_{i=1}^{m} e^{w_i}$$

7) Problème d'optimisation

On pose : $\varphi(\sigma,\theta) = \ln L(\sigma,\theta)$

Le problème consiste à rechercher les $n = s+2$ paramètres qui maximisent le log de la vraisemblance sous une contrainte de positivité : $\sigma > 0$.

$\varphi(\hat{\sigma},\hat{\theta}) \geq \varphi(\sigma,\theta),\ \forall (\sigma,\theta) \in \left(\mathbb{R}^{*+}\right) \times \mathbb{R}^{s+1}$

8) Dérivées premières et secondes :

Alors :

$$\begin{cases} \partial_{\theta_j}\varphi(\sigma,\theta) = \dfrac{1}{\sigma}\displaystyle\sum_{i=1}^{m} a_i . x_{ij} \\ \partial_{\sigma}\varphi(\sigma,\theta) = \dfrac{1}{\sigma}\displaystyle\sum_{i=1}^{m}\left[w_i . a_i - \delta_i\right] \end{cases} \quad \text{avec } a_i = e^{w_i} - \delta_i$$

pour les dérivées secondes :

$$\begin{cases} \dfrac{\partial^2\varphi(\sigma,\theta)}{\partial\sigma^2} = -\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{m}\left[{w_i}^2 . e^{w_i} + \delta_i\right] - \dfrac{2}{\sigma}\partial_{\sigma}\varphi(\sigma,\theta) \\ \dfrac{\partial^2\varphi(\sigma,\theta)}{\partial\sigma\partial\theta_j} = -\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{m} w_i . e^{w_i} . x_{ij} - \dfrac{1}{\sigma}\partial_{\theta_j}\varphi(\sigma,\theta) \\ \dfrac{\partial^2\varphi(\sigma,\theta)}{\partial\theta_j\partial\theta_k} = -\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{m} e^{w_i} . x_{ij} . x_{ik} \end{cases}$$

9) Algorithme de résolution :

La méthode de résolution proposée est une modification de la méthode de Levenberg-Marquardt pour tenir compte de la contrainte de positivité.

On cherche à minimiser $\psi = -\varphi$:

$$\psi(\hat{\sigma},\hat{\theta}) \le \psi(\sigma,\theta),\ \forall(\sigma,\theta) \in \left(\mathbb{R}^{*+}\right) \times \mathbb{R}^{s+1}$$

L'algorithme de Levenberg-Marquardt

La formule d'itération est la suivante :

$$\begin{pmatrix} \theta^{k+1} \\ \sigma^{k+1} \end{pmatrix} = \begin{pmatrix} \theta^{k} \\ \sigma^{k} \end{pmatrix} - \left(\mathrm{Hess}(\psi)(\theta^k,\sigma^k) + e_k\ {S_k}^2\right)^{-1} . \nabla\psi(\theta^k,\sigma^k) \qquad \textbf{(3)}$$

où $S_{ii}^2 = \mathrm{Hess}(\psi)(\theta^k,\sigma^k)(i,i) + \kappa$ et κ un nombre positif choisi pour assurer que S_{ii}^2 n'est pas nul ; par exemple $\kappa = 1$; et e_k une suite de réels positifs.

De façon pratique, e_k est choisi dans (3) pour que la matrice d'itération soit symétrique définie positive et suffisamment bien conditionnée, pour que numériquement la décomposition de Cholesky (cf. annexe) marche, et pour que l'on descende effectivement.

L'implémentation usuel de l'algorithme de Levenberg-Marquardt est la suivante :

On choisi deux paramètres Inc et Dec, tels que : $\text{Inc} > 1$ et $0 < \text{Dec} < 1$, par exemple, $\text{Inc} = 10$ et $\text{Dec} = 0{,}4$.

L'étape courante est :

Etant donné x^k et e_k,

a) assembler $\text{Hess}(\varphi)(\theta^k, \sigma^k)$ et $\nabla\varphi(\theta^k, \sigma^k)$;

b) faire une décomposition de Cholesky de $-\text{Hess}(\varphi)(\theta^k, \sigma^k) + e_k S_k{}^2$

Si (not.cholbon) faire $e_k \leftarrow \text{Inc}.e_k$ et recommencer la décomposition (ce processus est nécessairement fini) ;

c) puis Calculer $\theta^{k+1}, \sigma^{k+1}$ et $\varphi(\theta^{k+1}, \sigma^{k+1})$;

si $\left(\left\| \begin{pmatrix} \theta^{k+1} \\ \sigma^{k+1} \end{pmatrix} - \begin{pmatrix} \theta^k \\ \sigma^k \end{pmatrix} \right\| \le \varepsilon . \left\| \begin{pmatrix} \theta^k \\ \sigma^k \end{pmatrix} \right\| \right)$ stop ;

si $\left(\varphi(\theta^{k+1}, \sigma^{k+1}) \le \varphi(\theta^k, \sigma^k) \right)$ faire $e_k \leftarrow \text{Inc}.e_k$ et retour en b) ;

si $\left(\varphi(\theta^{k+1}, \sigma^{k+1}) > \phi(\theta^k, \sigma^k) \right)$ faire $e_k \leftarrow \text{Dec}.e_k$.

La prise en compte de la contrainte se fait en modifiant la première ligne de c) :

c') Calculer $\theta^{k+1}, \sigma^{k+1}$;

si $((\sigma^{k+1} > 0))$ calcul de $\varphi(\theta^{k+1}, \sigma^{k+1})$;

si $\left(\left\| \begin{pmatrix} \theta^{k+1} \\ \sigma^{k+1} \end{pmatrix} - \begin{pmatrix} \theta^k \\ \sigma^k \end{pmatrix} \right\| \le \varepsilon . \left\| \begin{pmatrix} \theta^k \\ \sigma^k \end{pmatrix} \right\| \right)$ stop ;

si $\left(\varphi(\theta^{k+1}, \sigma^{k+1}) \le \varphi(\theta^k, \sigma^k) \right)$ ou $\sigma^{k+1} \le 0$ faire $e_k \leftarrow \text{Inc}.e_k$ et retour en b).

Notons $(\hat{\theta}, \hat{\sigma})$ l'estimation de l'optimum local : (θ, σ), vers lequel a convergé l'algorithme.

matrice de variance covariance

La matrice de variance-covariance est estimée comme :

$$V = -\left(\text{Hess}(\varphi)(\hat{\theta}, \hat{\sigma})\right)^{-1}$$

test de Wald

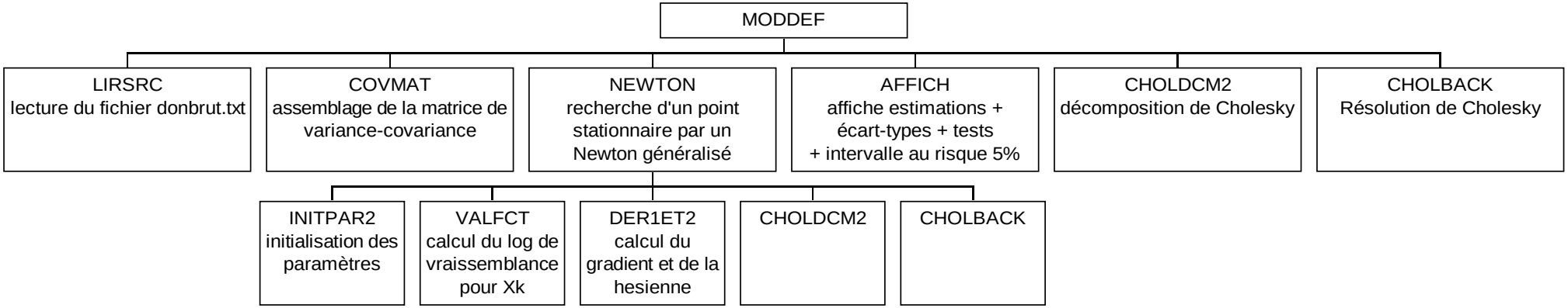
Pour tester si un une composante de θ est significativement différents de 0, on calcule :

$$w_i = \frac{\hat{\theta}_i^2}{V_{ii}}$$

qui suit sous l'hypothèse nulle $H_0 : \theta_i = 0$ une loi de χ_1^2.

de la même façon : $w_n = \frac{(\hat{\sigma} - 1)^2}{V_{nn}}$ suit un χ_1^2.

Arborescence



Non listés (uniquement PC) :

attente : permet de garder à l'écran un affichage en attendant une entrée clavier

posicurs : positionne le curseur

clrscr : efface l'écran

Fonctions

epsil (real*4) : un epsilon machine / $1+\varepsilon=1$.

zegal (logical) : permet de comparer deux réels*8.

DESCRIPTION DU FICHIER DE DONNEE LU PAR LE SOUS-PROGRAMME LIRSRC : DONBRUT.TXT

```
---------1----1----2----2----3----3----4----4----5----5----6---
----5----0----5----0----5----0----5----0----5----0----5----0---
```

Premier d'enregistrement :

format d'origine : I5,2(1x,I2),1X,f5.2,1X,I2

```
  a   b  c    d   e
1710  5  1   .10  1
```

a) nombre d'individus
b) nombre de covariables
c) nombre de covariables qualitatives
d) seuil permettant de tester si Weibull stratifié ou pas.
e) niveau de sévérité pour affichage à l'exécution
   0 aucun affichage
   1 affichage intermédiaire
   2 mode bavard

deuxième d'enregistrement :

format d'origine : 10(1x,I2)

```
a  b  c  d  e
1  1  1  2  1
```

a) b) c) d) et e) ici pour chacune des covariables ici au nombre de cinq :
   1 si elle doit être traitée en variable continue
   0 si on ne doit pas en tenir compte
   >1 le nombre de modalité de cette variable qualitative

troisième enregistrement :

format d'origine : 10(1x,A10)

```
 a             b         c          d        e
sol        logdiam/60 loglon/500 matériau   trafcha
```

a) b) c) d) et e) nom de chacune des covariables

Les quatrièmes et cinquièmes enregistrement se répète autant de fois qu'il y a de variables qualitatives.

quatrième type d'enregistrement:

format d'origine : 10(1x,f6.2)

```
  a       b
1.00    .00
```

a) b) valeurs de chaque modalité

cinquième type d'enregistrement :

format d'origine : 10(1x,A10)

```
  a             b
fonte duct fonte gris
```

a) b) noms de chaque modalité

sixième type d'enregistrement :

format d'origine : 1x,I5,2(1x,I2),1x,I1,2(1x,I2),10(1x,f7.2)

```
a  b c  d e   f     g       h       i       j       k
1  1 91 0 27  0     .00     .00    -.93     .00     .00
```

a) identifiant tronçon
b) code commune
c) date de fin d'observation ou de défaillance
d) 0 si c) est une date de fin d'observation 1 sinon
e) durée entre date de dernière défaillance ou date de pose et date donnée en c). C'est la variable d'intérêt
f) Nombre de défaillances antérieures
G) h) i) j) et k) valeurs de chacune des covariables.

DESCRIPTION DU FICHIER PARAMETRES ESTIMES : PARESTI.TXT

Hypothèses n°1 : trois familles de conduites identifiées dans l'étude (cf. annexe technique)

```
NBA  5 nombre de paramètres de régression
ND1   .03  1.73            première strate : λ1 et p1
ND2   .14  1.06            deuxième strate : λ2 et p2
ND3   .22  1.46           troisième strate : λ3 et p3
sol           .422    .131    .711
```

```
logdiam/60   -.918   -.863  -1.517
loglon/500    .683    .541    .516
fonte duct   -.198    .064    .918
fonte gris    .000    .000    .000
trafcha       .847    .192   -.121
```

Hypothèses n°2 : deux familles de conduites identifiées dans l'étude (cf. annexe technique) - existence sur le réseau d'un nombre très restreint de conduites ayant subi plus de trois casses. Ici même paramètres de régression pour les deux strates.

```
NBA  5 nombre de paramètres de régression
   .03  1.73            première strate : λ1 et p1
   .17  1.07            deuxième strate : λ2 et p2
sol           .352
logdiam/60  -1.010
loglon/500    .610
fonte duct    .041
fonte gris    .000
trafcha       .324
```

DESCRIPTION DU FICHIER PREVISION DES DEFAILLANCES : DEFPREV.TXT

```
---------1----1----2----2----3----3----4----4----5----5----6---
----5----0----5----0----5----0----5----0----5----0----5----0---
```

format d'origine : 1x,I5,2(1X,I2),1X,I1,2(1X,I2),1x,f10.4

```
    a   b c  d  e  f        g
  424   6 92 0  4  4     7.3372
  417   6 92 0  1  5     2.8474
  196   6 92 0  5  5     1.6197
  540   6 92 0  8  5     4.4441
```

423	6	92	0	5	5	6.8159
24	6	92	0	7	5	6.3252
4	6	92	0	9	5	3.8046
370	6	92	0	4	5	1.6093

a) identifiant tronçon
b) code commune
c) date de fin d’observation ou de défaillance
d) 0 si c) est une date de fin d’observation 1 sinon
e) durée entre date de dernière défaillance ou date de pose et date donnée en c). C’est la variable d’intérêt
f) Nombre de défaillances antérieures
G) Nombre de défaillances prévues.