Modélisation d'événements récurrents sous hypothèse d'accumulation de défaillances (Worse than old assumption)

G. Babykina, V. Couallier, Y. Le Gat

Cemagref, Université de Bordeaux

Novembre 19, 2009

UNIVERSITÉ DE BORDEAUX

Cemagref
Sciences, eaux & territoires

Plan

Données à analyser

- Défaillances des canalisations d'un réseaux d'eau potable (individu : tronçon, événement : casse).
- Beaucoup de données sur longues périodes.
- Une casse entraîne le remplacement (fin de vie) ou la réparation ⇒ 1 type de maintenance (corrective).
- Pour certaines vieux matériaux : censure à gauche.
- Enjeu : optimiser la politique de remplacement.

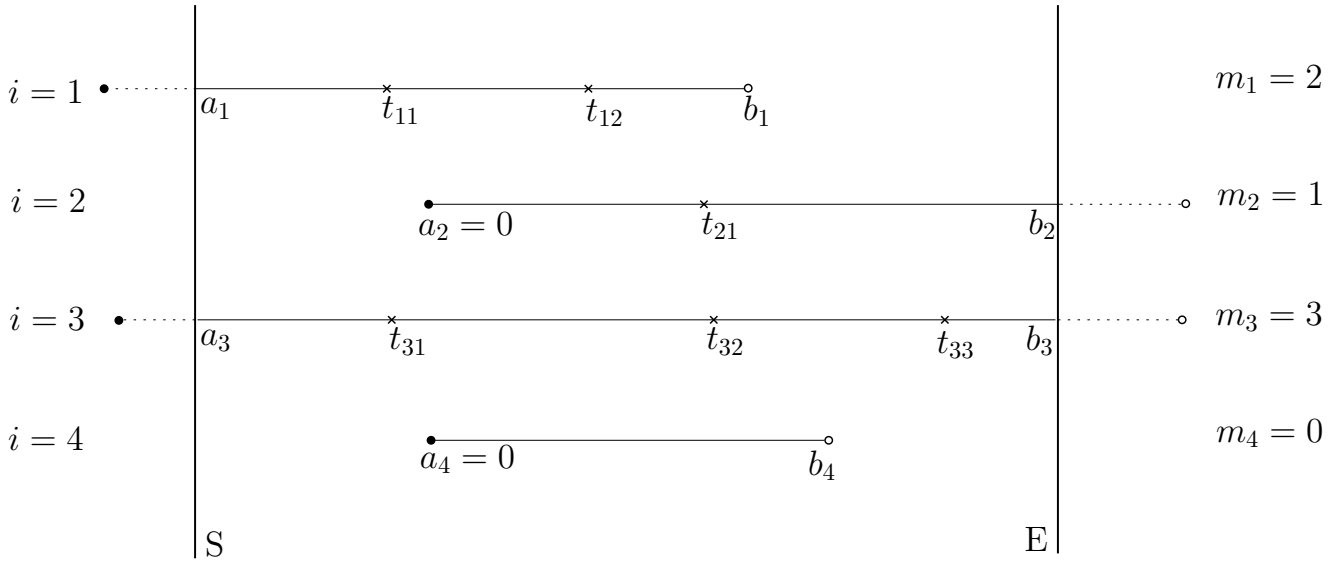
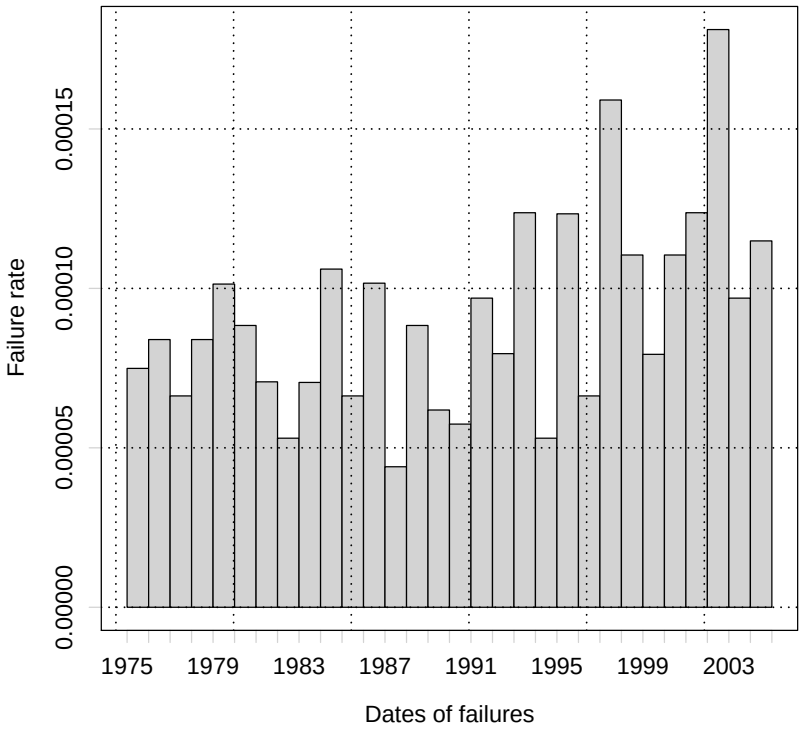
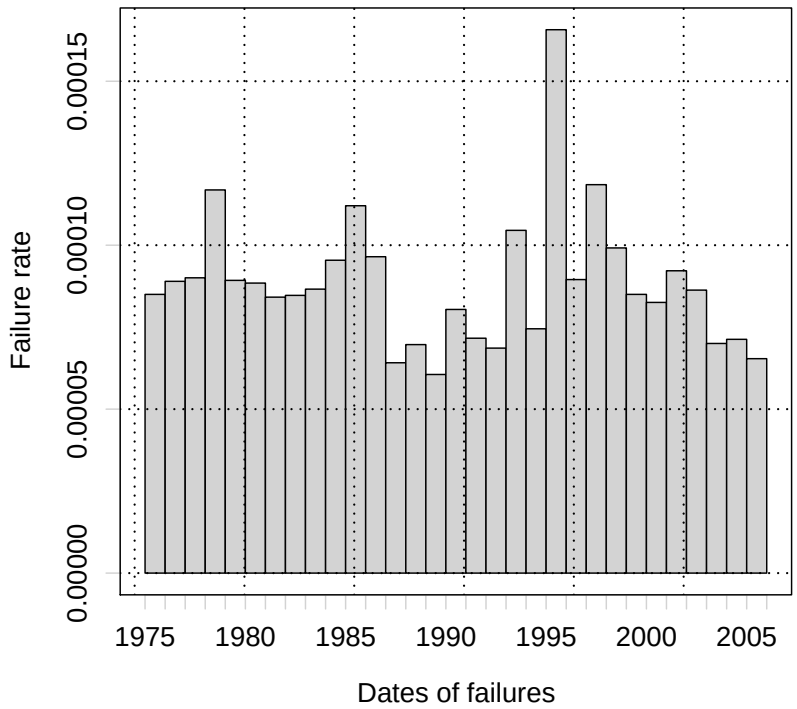


Illustration (1/2)

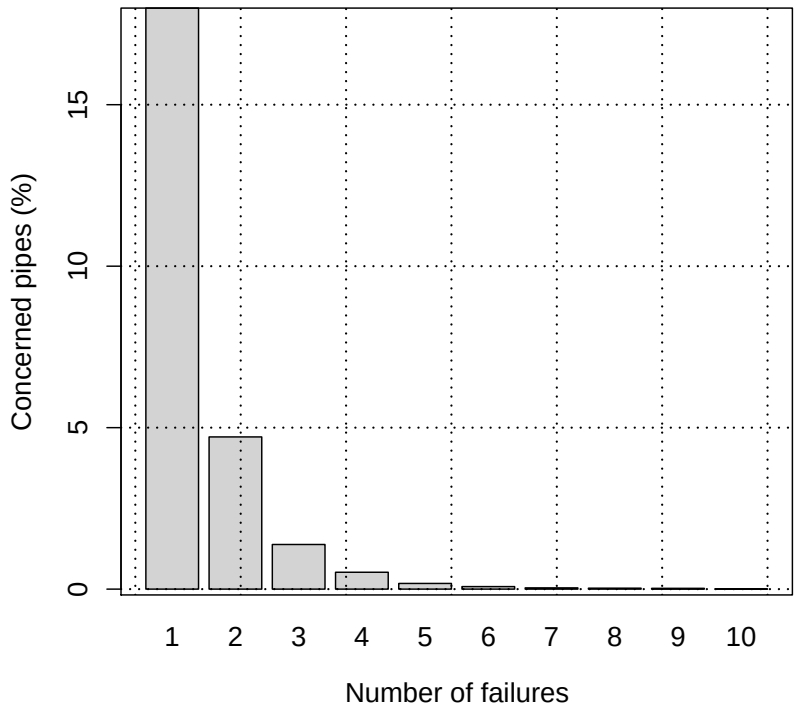


Vue globale : vieillessement.



Phénomènes temporels.

Illustration (2/2)



Accumulation des défaillances.

- Fragilité initiale individuelle ? (défauts d'installation).
- Effet nuisible des maintenances (Worse than Old) ?

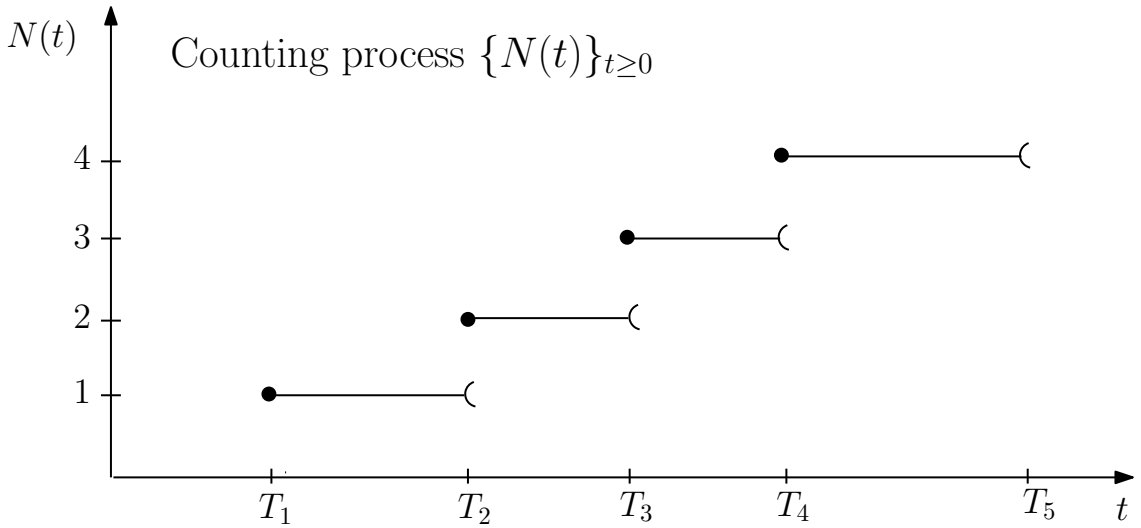
Phénomènes à considérer

- Censure à gauche.
- Censure de type I à droite.
- Maintenances correctives.
- Vieillessement.
- Accumulation des défaillances.
- Covariables (fixes et dépendant du temps).

Plan

Processus de comptage

- Processus de comptage [Andersen et al., 1993].



- Intensité instantanée :

$$\begin{aligned}\lambda(t) &= \mathbb{E}\left[dN(t) \,|\, \mathcal{H}(t)\right] \\ &= \lim_{dt\to 0} \frac{1}{dt}\mathbb{P}\left[N(t+dt) - N(t) = 1 \,|\, \mathcal{H}(t)\right]\end{aligned}$$

avec $\mathcal{H}(t) = \sigma\left(N(t), T_1, ..., T_{N(t)}, Z(t)\right)$.

LEYP : Linear Extension to Yule Process

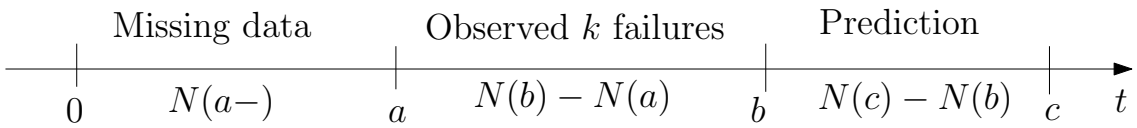
- LEYP [Le Gat, 2009] : NHPP + processus de Yule + Cox PH :

$$\lambda(t) = (1 + \alpha j)\delta t^{\delta-1} e^{Z(t)'\beta}\, dt$$

- Accumulation des défaillance ($\alpha > 0$, j : nombre de défaillances précédentes).
- Vieillessement ($\delta > 1$, t : l'âge d'individu).
- Hétérogénéité individuelle, phénomènes temporels ($Z(t)$: vecteur de covariables).
- Paramètres à estimer : $\theta = \{\alpha, \delta, \beta\}$
- Notations :
  - Intensité initiale : $\lambda_0(t) = \delta t^{\delta-1} e^{Z(t)'\beta}\, dt$.
  - Intensité initiale cumulée : $\Lambda_0(s) = \int_0^s \delta u^{\delta-1} e^{Z(u)'\beta}\, du$.
  - $Z(t)'\beta = \beta_0 + \beta_1 Z_1 + ... + Z_p(t)\beta_p$.

Distribution Binomiale Négative

- Processus de Yule (processus de naissance) [Ross, 1983] : $\mathbb{E}[dN(t) \,|\, N(t-) = j] = j\lambda dt$, distribution de $N(t)$ sur k fondateur est Binomiale Négative.
- LEYP : $\mathbb{E}[dN(t) \,|\, N(t-) = j] = (1 + \alpha j)\lambda_0(t)\, dt$.



$$[N(c) - N(b) \,|\, N(b-) - N(a) = k] \sim \mathcal{NB}\left(\alpha^{-1} + k, \frac{\mathrm{e}^{\alpha\Lambda_0(b)} - \mathrm{e}^{\alpha\Lambda_0(a)} + 1}{\mathrm{e}^{\alpha\Lambda_0(c)} - \mathrm{e}^{\alpha\Lambda_0(a)} + 1}\right)$$

- $\Rightarrow$ Prise en compte de la censure à gauche $[N(a-)|N(b) - N(a)]$.
- $\Rightarrow$ Prédiction de futures défaillances $[N(c) - N(b)|N(b) - N(a)]$.

Vraisemblance

- Vraisemblance pour 1 individu avec m événements aux moments t_j ($j = \{1, \ldots m\}$) (forme générale) :

$$L(\theta) = \left(\prod_{j=1}^{m} \lambda(t_j) \right) \times \exp \left(- \sum_{j=0}^{m} \int_{t_j}^{t_{(j+1)}} \lambda(u)\, du \right)$$

- LEYP log-vraisemblance pour 1 tronçon avec m défaillances pendant la période $[a, b]$:

$$\begin{aligned} \ln L(\theta) &= m \ln \alpha + \ln \Gamma(\alpha^{-1} + m) - \ln \Gamma(\alpha^{-1}) \\ &- (\alpha^{-1} + m) \ln(e^{\alpha \Lambda_0(b)} - e^{\alpha \Lambda_0(a)} + 1) \\ &+ \sum_{j=1}^{m} (\ln \lambda_0(t_j) + \alpha \Lambda_0(t_j)) \end{aligned}$$

  - $\lambda_0(t) = \delta t^{\delta - 1} e^{Z(t)' \beta}\, dt$

Cadre théorique général

"A general class of parametric models for recurrent event data" [Peña, 2006], [Stocker and Peña, 2007].

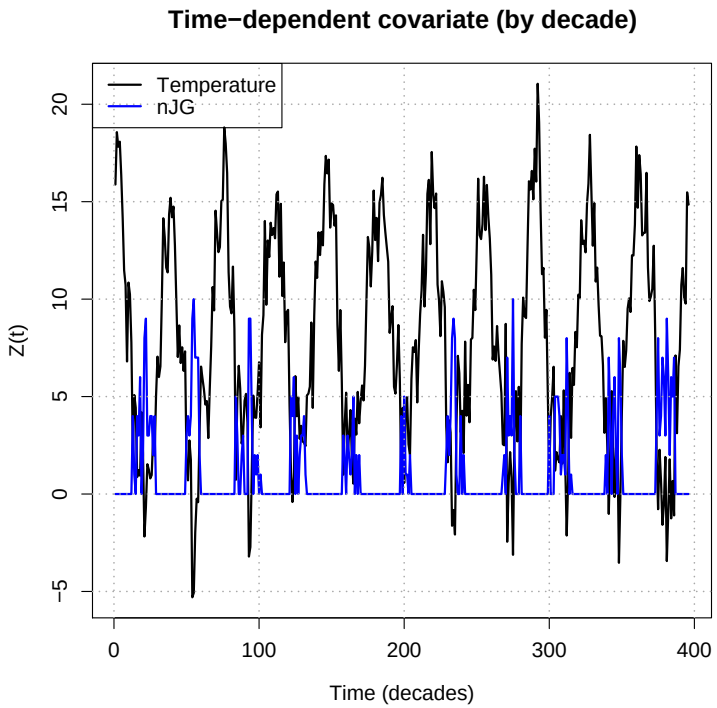
$$\lambda(t;\, Z, X) = Z\lambda_0[\varepsilon(t);\theta]\rho[N(t-);\alpha]\psi[\beta' X]$$

- Z : fragilité individuelle.
- $\lambda_0[\cdot\ ;\ \theta]$: intensité initiale, fonction de l'âge (virtuel).
- $\rho[\cdot\ ;\ \alpha]$: fonction d'accumulation de défaillances.
- $\psi[\cdot\ ;\ \beta]$: fonction des covariables.

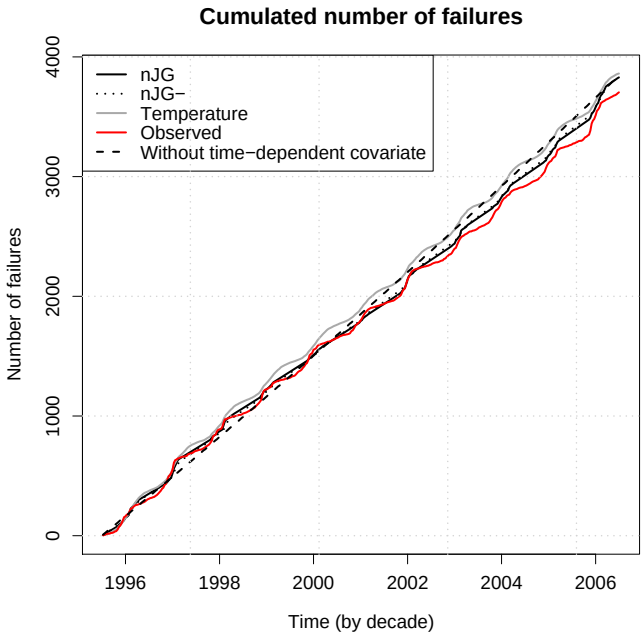
Plan

Exemple sur les données réelles (1/2)

- Données : 21450 tronçons, 3704 défaillances, 10 ans d'observation.
- Modèle : $\lambda(t) = (1 + \alpha N(t-))\delta t^{\delta-1} e^{\beta_0+\beta_1 Z_1+\beta_2 Z_2(t)}$.
- Covariables :
  - Z_1 : covariable fixe (logarithme de la longueur du tronçon) ;
  - $Z_2(t)$: covariable temporelle, (le nombre de jours de gel (nJG) et la température moyenne ($Temp$) mesurés par décade).



Exemple sur les données réelles (2/2)



$Z_2(t)$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	Erreur$^{(1)}$
nJG	0.96	1.15	-8.09	0.69	0.18	3.37%
$nJG-^{(2)}$	0.96	1.13	-7.88	0.68	0.15	3.45%
$Temp$	1.02	0.91	-5.81	0.54	-0.10	4.24%
$-^{(3)}$	0.98	1.11	-7.52	0.65	–	

$^{(1)}$Erreur de "prédiction" (casses observées vs. casses prédites par le modèle sur la période d'observation).
$^{(2)}$Nombre de jours du gel de la décade précédente.
$^{(3)}$Estimations et "prédictions" sans la covariable temporelle.

- Choix de $Z_2(t)_{t\leq a}$ dans le calcul de $\Lambda_0(a) = \int_0^a \lambda_0(s)\,ds$ dépend de la nature de $Z_2(t)$.
  - $Z_2(t)_{t\leq a} = 0$;
  - $Z_2(t)_{t\leq a} = \overline{Z}_2(t)_{a\leq t\leq b}$ (moyenne sur la période d'observation) ;
  - 2 constantes :
    $\lambda_0(t) = \left(\delta t^{\delta-1} \mathrm{e}^{\beta_0^1+\beta_1 Z_1+\beta_2 Z_2(t)}\right) \mathbb{1}_{t\geq a} + \left(\delta t^{\delta-1} \mathrm{e}^{\beta_0^2+\beta_1 Z_1}\right) \mathbb{1}_{t\leq a}$

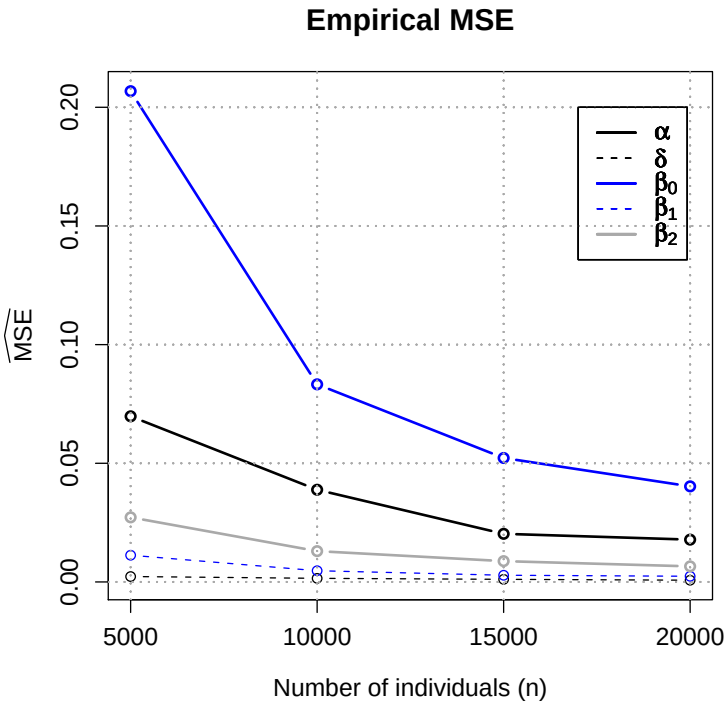
Simulations : cadre général

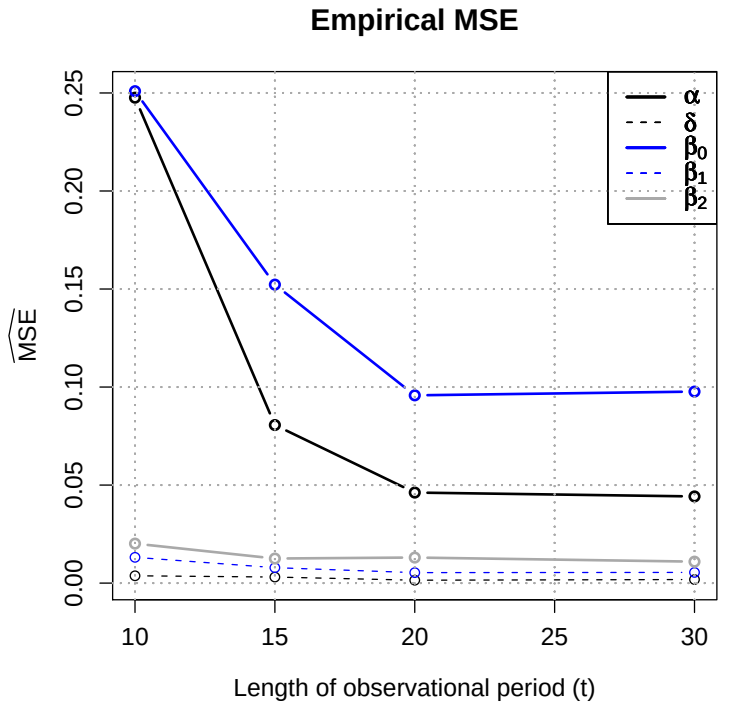
- Les cas considérés :
  1. Nombre d'individus $n = \{5000, 10000, 15000, 20000\}$, période d'observation $t = 30$ ans.
  2. Période d'observation $([a, b])$ $t = \{10, 15, 20, 30\}$ ans, nombre d'individus $n = 10000$.
- 100 simulations.
- Pose $[(a - 20), b]$.
- Covariable fixe, $Z_1 \sim \varepsilon(0.01)$ (imitation de la longueur d'un tronçon).
- Covariable temporelle $Z_2 = \mathbb{1}_{(b-5)}$: année "à risque".
- Le modèle :

$$\mathbb{E}[dN(t) \mid N(t-) = j] = (1 + \alpha j)\delta t^{\delta-1} \mathrm{e}^{\beta_0+\beta_1 Z_1+\beta_2 Z_2(t)}\, dt$$

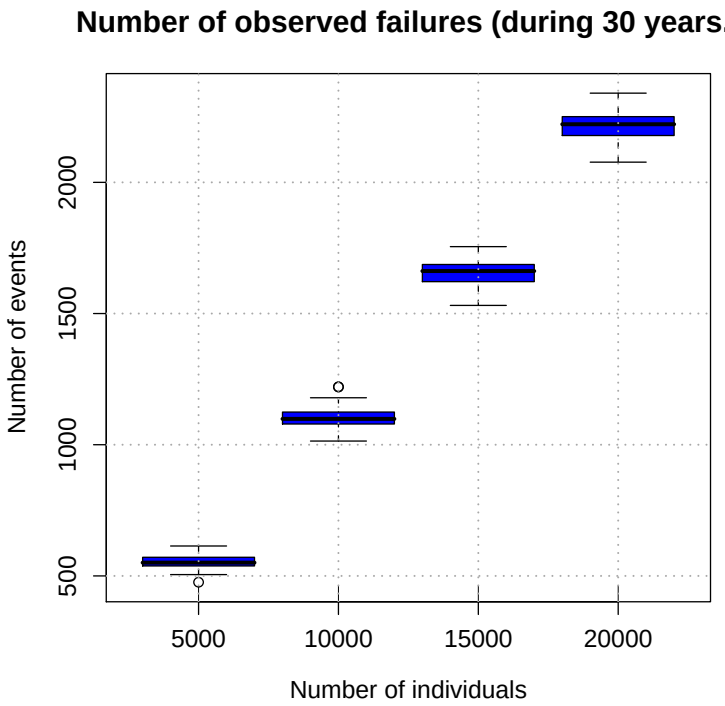
- Vecteur de paramètres à estimer $\theta = \{\alpha, \delta, \beta_0, \beta_1, \beta_2\}$.
- Paramètres d'évaluation : MSE empirique ($\widehat{MSE}$), biais relatif $\frac{|\hat{\theta}_k - \theta_k|}{\theta_k}$, coefficient de variation (CV) de $\hat{\theta}_k$.

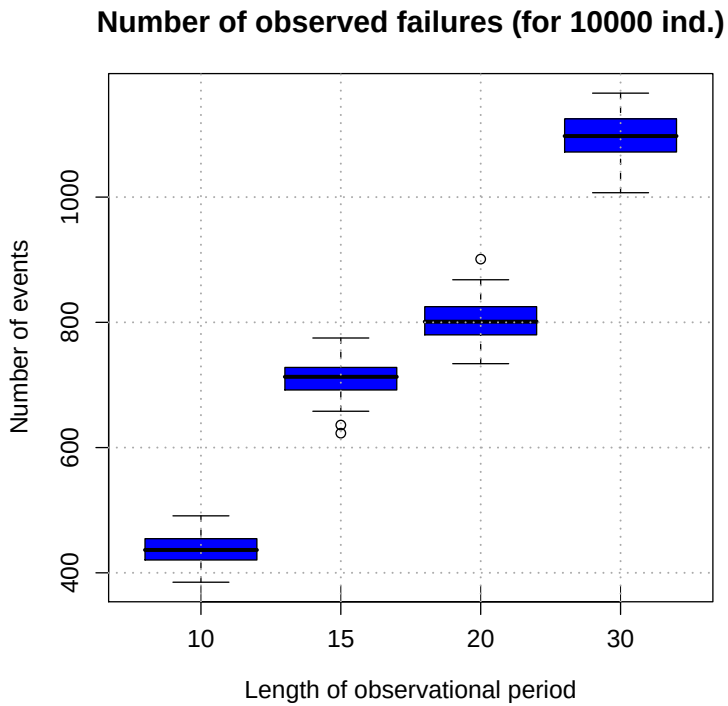
MSE empirique



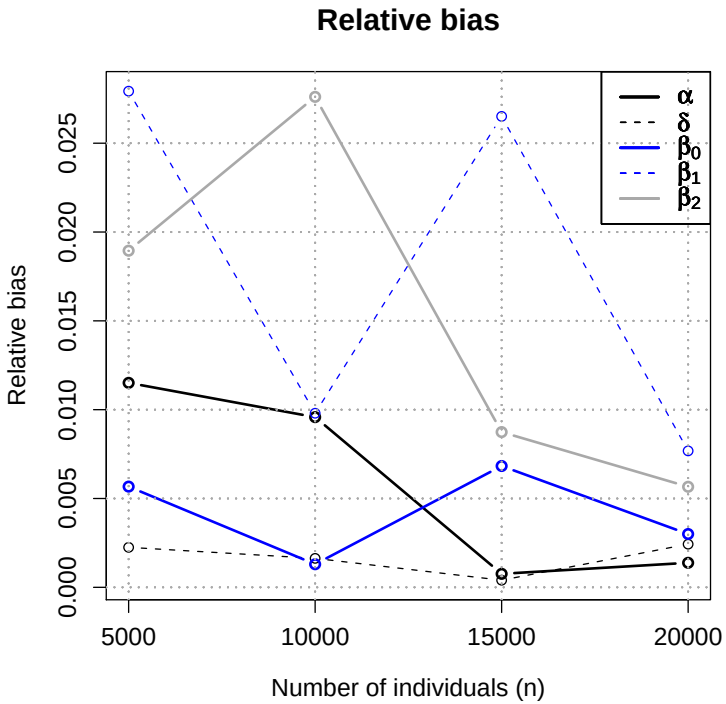


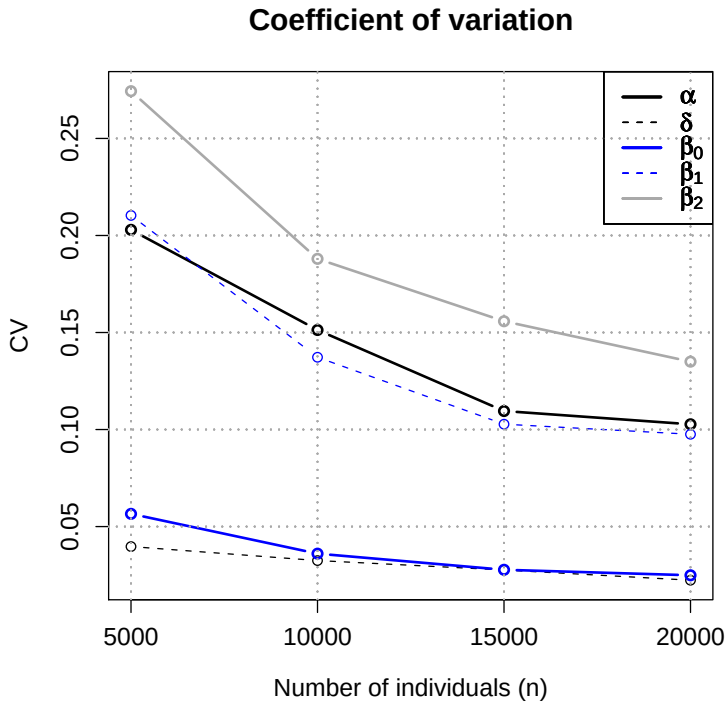
Nombre de défaillances (quantité d'information)



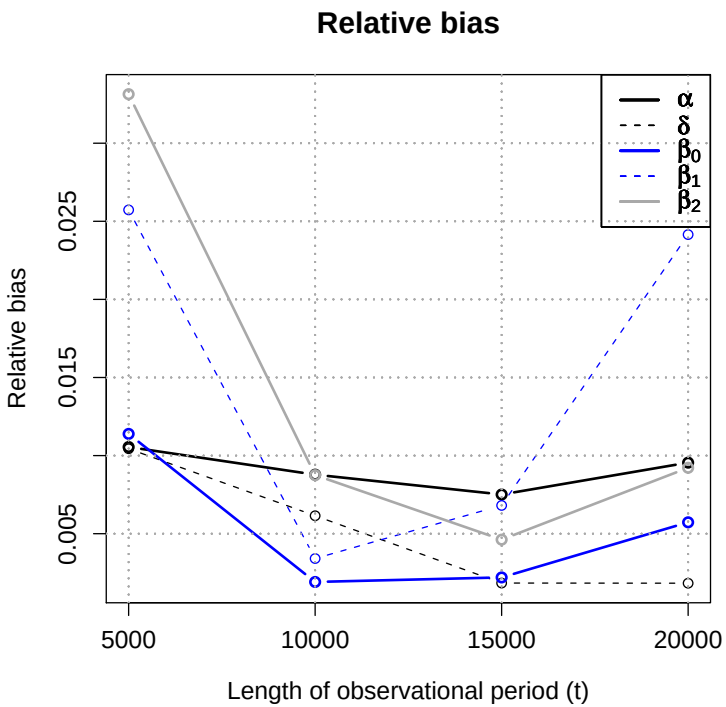


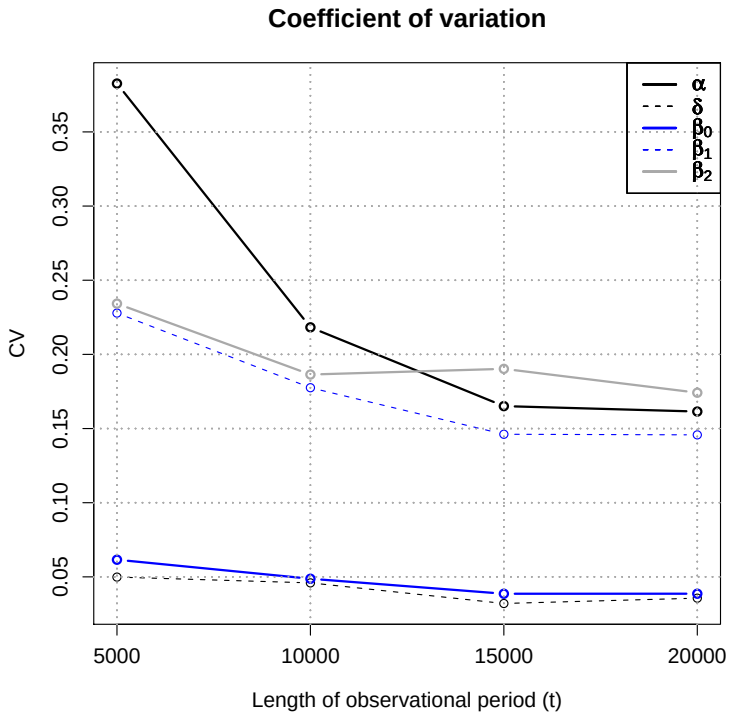
Biais relatif et CV vs. nombre d'individus





Biais relatif et CV vs. longueur de période d'observation





LEYP : récapitulatif

- On maîtrise :
  - estimation des paramètres (simulations et données réelles) ;
  - intégration d'une covariable temporelle (sous différentes formes) ;
  - propriétés "asymptotiques" empiriquement ;
  - évaluation de qualité d'ajustement par prédictions (validation croisée).
- On ne maîtrise pas :
  - propriétés double-asymptotiques (temps et individus) ;
  - propriétés asymptotiques théoriques ;
  - tests formels de la qualité d'ajustement ;
  - ....

Plan

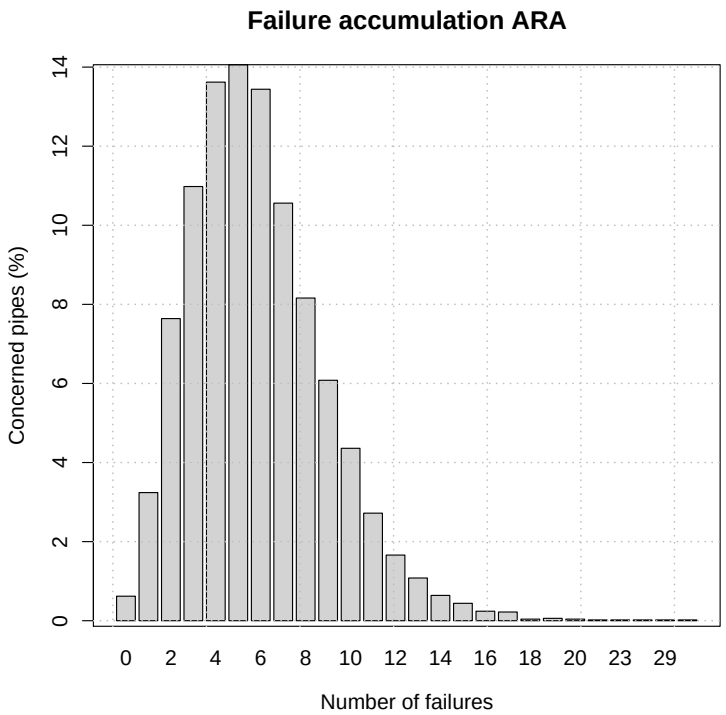
Les modèles : ARA_∞ vs. LEYP

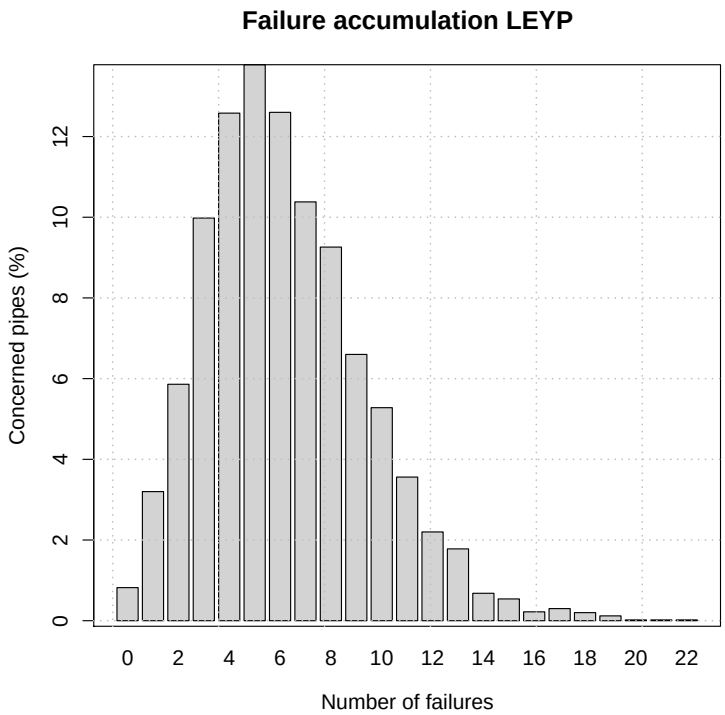
- Modèle ARA_∞ [Doyen and Gaudoin, 2004], [Corset et al., 2009].

$$\begin{cases} \lambda_t = \lambda\left(t - \rho\sum_{j=0}^{N_t-1}(1-\rho)^j T_{N_t-j}\right), & N_t > 0 \\ \lambda(t) = \alpha\beta t^{\beta-1} \end{cases}$$

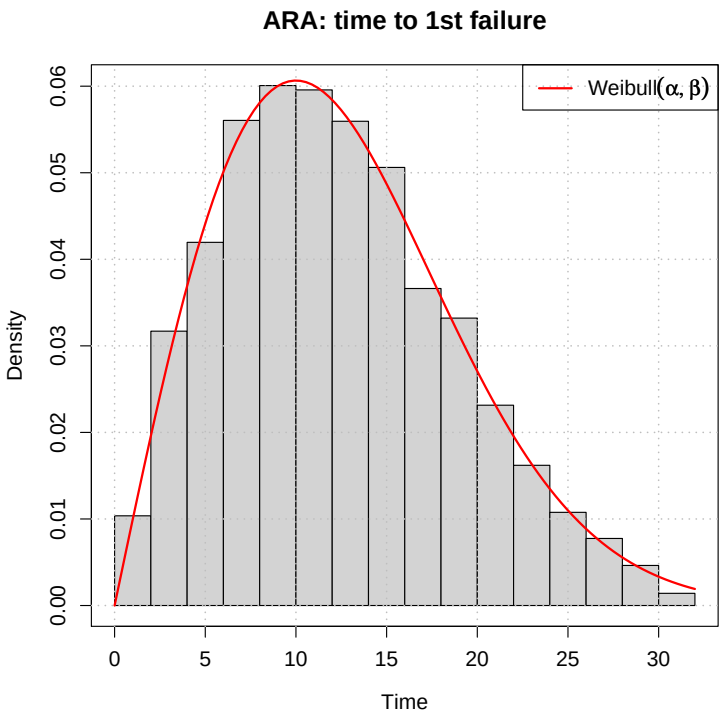
- "Equivalence" des paramètres :
  - $\alpha^{\text{ARA}} \equiv \left(\mathrm{e}^{\beta_0}\right)^{\text{LEYP}}$ (constante) ;
  - $\beta^{\text{ARA}} \equiv \delta^{\text{LEYP}}$ (vieillessement) ;
  - $\rho^{\text{ARA}} \equiv \alpha^{\text{LEYP}}$ (accumulation des défaillances, effet de maintenance).
- Exemple.
  - Données simulées selon ARA_∞ avec $\alpha = 0.005$, $\beta = 2$, $\rho = -0.1$.
  - 5000 individus. Sans covariables.
  - Paramètres estimés avec LEYP : $\mathrm{e}^{\hat{\beta}_0} = 0.0051$, $\hat{\delta} = 1.98$, $\hat{\alpha} = 0.099$.
  - Voir le graphe d'intensités.

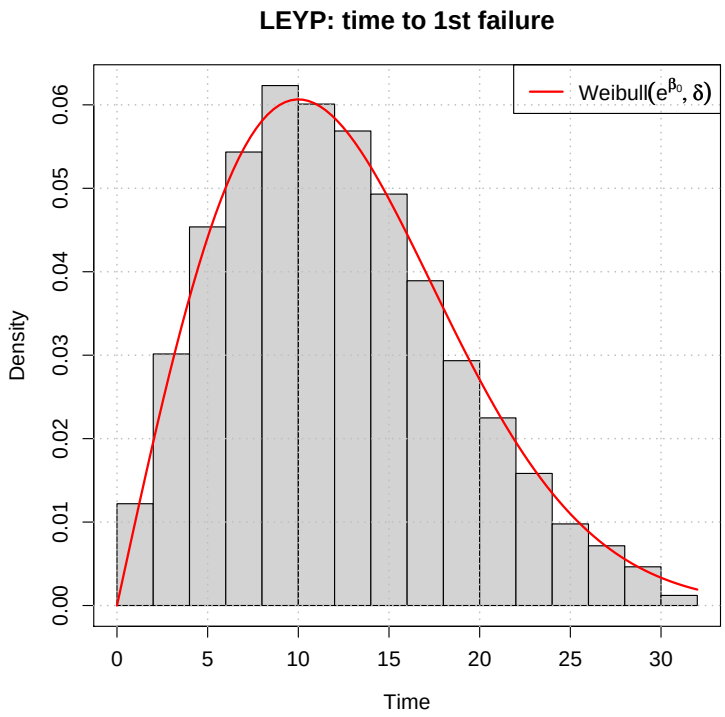
Accumulation des défaillances





Temps jusqu'à première défaillance





	Maximum défaillances	Défaillances totale	Défaillances par individu
ARA_∞	30	29331	5.9
LEYP	22	31035	6.2

LEYP estimé avec ARA$_\infty$: résultats (1/2)

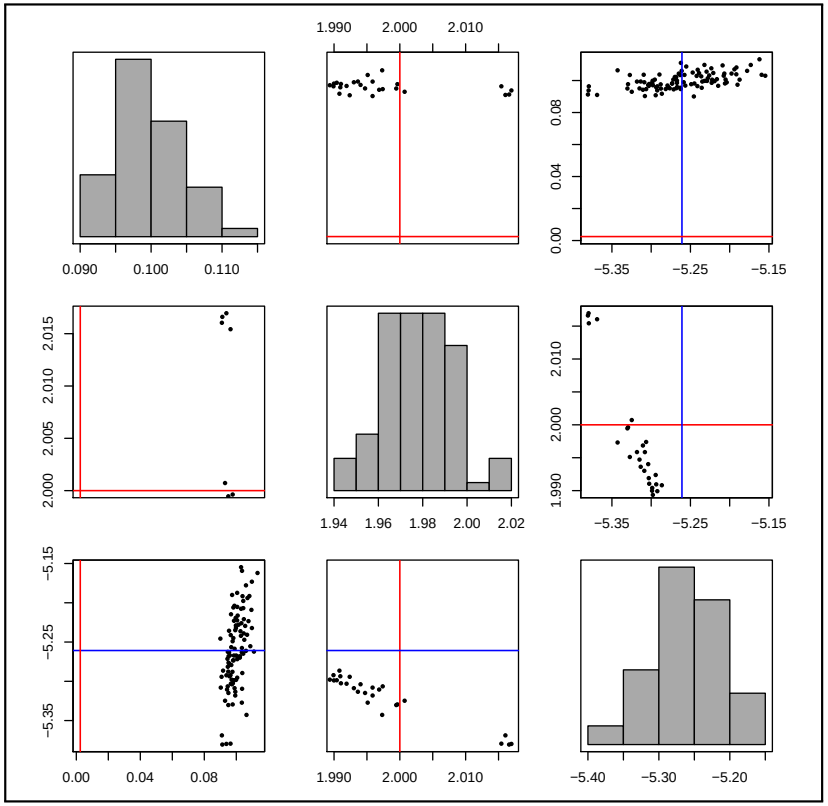
1. Jeu de données 1 (100 simulations MC).
   - Simulé : LEYP avec $\theta = \{1.5, 2, \ln(0.0025)\}$, 100 individus, 30 ans d'observation. En moyenne : 23.5 défaillances/individu, 154.2 max défaillances/individu.
   - Paramètres d'ARA$_\infty$ estimés :

	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\rho}$
Moyenne	0.00318	2.45986	-0.0278
SD	2e-04	0.0459	0.007

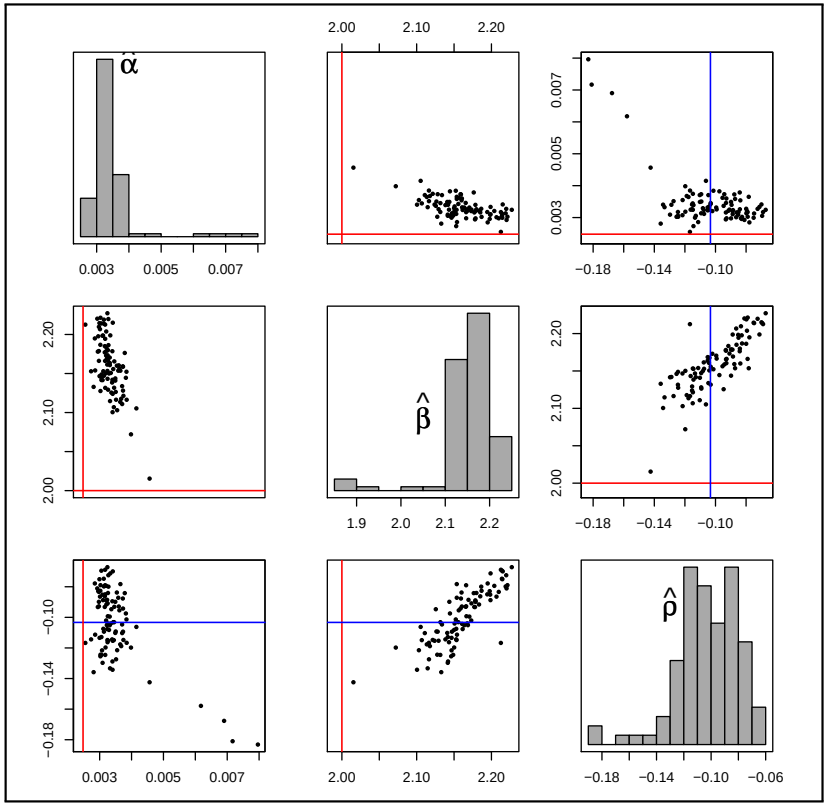
2. Jeu de données 2 (100 simulations MC).
   - Simulé : LEYP avec $\theta = \{0.9, 2, \ln(0.0025)\}$, 100 individus, 30 ans d'observation. En moyenne : 8.5 défaillances/individu, 42.9 max défaillances/individu.
   - Paramètres d'ARA$_\infty$ estimés :

	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\rho}$
Moyenne	0.00331	2.15944	-0.10043
SD	3e-04	0.0364	0.0179

LEYP estimé avec ARA_∞ : résultats (2/2)



LEYP : $\alpha = 1.5$, $\beta = 2$,
$e^{\beta_0} = 0.0025$.



LEYP : $\alpha = 0.9$, $\beta = 2$,
$e^{\beta_0} = 0.0025$.

- Optimisation sous contrainte $\rho < 0$, problèmes de convergence.
- En rouge : "vraies" valeurs, en bleu : moyenne empirique.

ARA_∞ estimé avec LEYP : résultats (1/2)

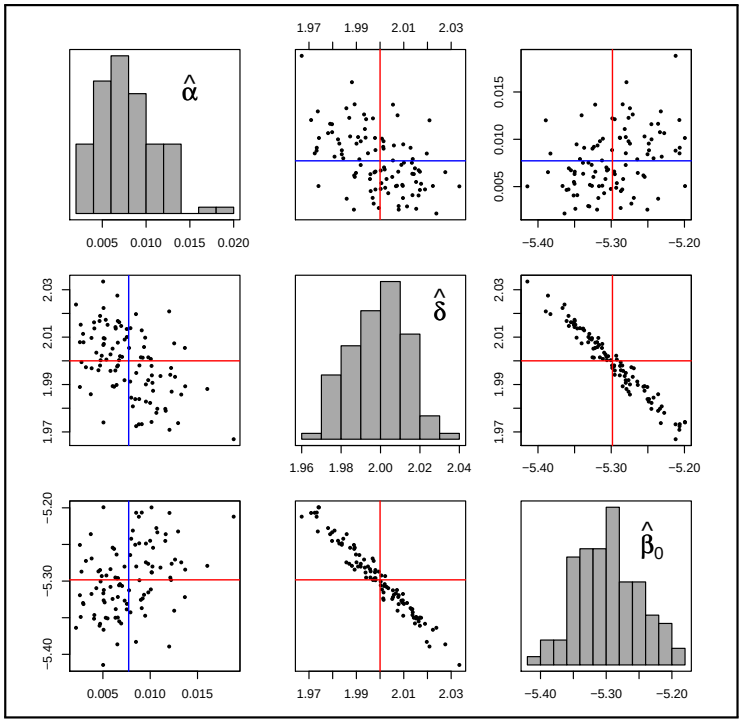
1. Jeu de données 1 (100 simulations MC).
   - Jeu de donnée simulé : ARA_∞ avec $\theta = \{e^{-5.3}, 2, -0.01\}$, 5000 individus, 30 ans d'observation. En moyenne : 4.88 défaillances/individu, 15 max défaillances/individu.
   - Estimé avec LEYP (sur 100 estimations) :

	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\beta}_0$
Moyenne	0.00772	1.99898	-5.29725
SD	0.0032	0.0143	0.0471

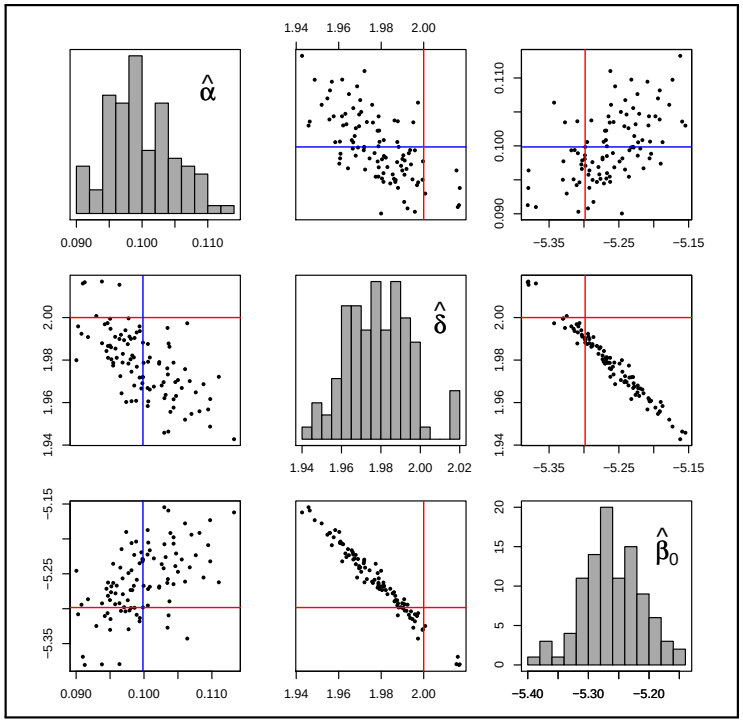
2. Jeu de donnée 2 (100 simulations MC).
   - Simulé : ARA avec $\theta = \{e^{-5.3}, 2, -0.1\}$, 5000 individus, 30 ans d'observation. En moyenne : 5.89 défaillances/individu, 25.9 max défaillances/individu.
   - Estimé avec ARA_∞ :

	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\beta}_0$
Moyenne	0.09984	1.9781	-5.26081
SD	0.005	0.0158	0.0488

ARA_∞ estimé avec LEYP : résultats (2/2)



ARA_∞ : $\alpha = 0.005$, $\beta = 2$, $\rho = -0.01$.



ARA_∞ : $\alpha = 0.005$, $\beta = 2$, $\rho = -0.1$.

- En rouge : "vraies" valeurs de paramètres, en bleu : moyenne empirique sur 100 estimations.

Plan

Martingales : définition, propriétés

- $M(t) = N(t) - A(t) = N(t) - \int_0^t \mathbb{E}\left[dN(s) \,|\, \mathcal{H}(s-)\right]$ est une martingale par rapport à $\mathcal{H}(s-)$.
- Propriétés d'une martingale :
  1. $\mathbb{E}\left[dM(t) \,|\, \mathcal{H}(t-)\right] = 0$ (centrée).
  2. $\mathrm{Cov}\left[M(t),\, M(t+u) - M(t)\right] = 0 \quad u, t > 0$ (incréments non-corrélés).
- Le résidu martingale pour l'individu i à chaque défaillance :

$$\widehat{M}_i(t_{i,j}) = N_i(t_{i,j}) - \int_0^{t_{ij}} (1 + \hat{\alpha} N_i(u)) \hat{\delta} u^{\hat{\delta}-1} \mathrm{e}^{Z_i'(u)\hat{\beta}} du, \quad j = 1 \ldots m_i$$

- Incréments du compensateur et du résidu martingale entre 2 événements :

$$\Delta\widehat{M}_i(t_{i,j}) = \widehat{M}_i(t_{i,j}) - \widehat{M}_i(t_{i,(j-1)})$$

$$\Delta\widehat{A}_i(t_{i,j}) = \widehat{A}_i(t_{i,j}) - \widehat{A}_i(t_{i,(j-1)})$$

Un test d'ajustement (1/2)

- "Omnibus Tests of The Martingale Assumption in The Analysis of Recurrent Failure Time Data" [Jones and Harrington, 2001].
- Modèle semi-paramétrique.
- J : nombre d'événements pré-défini (maximal).
- $\widehat{M}_i(t_{i,0}) = 0$. $\widehat{M}_i(t_{i,j}) = 0$ après censure.
- Matrice D de dimension $n \times J$:

$$D = \begin{pmatrix} \Delta\widehat{M}_1(t_{1,1}) & \cdots & \Delta\widehat{M}_1(t_{1,J}) \\ \vdots & \ddots & \vdots \\ \Delta\widehat{M}_n(t_{n,1}) & \cdots & \Delta\widehat{M}_n(t_{n,J}) \end{pmatrix}$$

- μ : vecteur J-dimensionnel des moyennes par colonne théorique estimé par $\overline{\Delta M}$.

Un test d'ajustement (2/2)

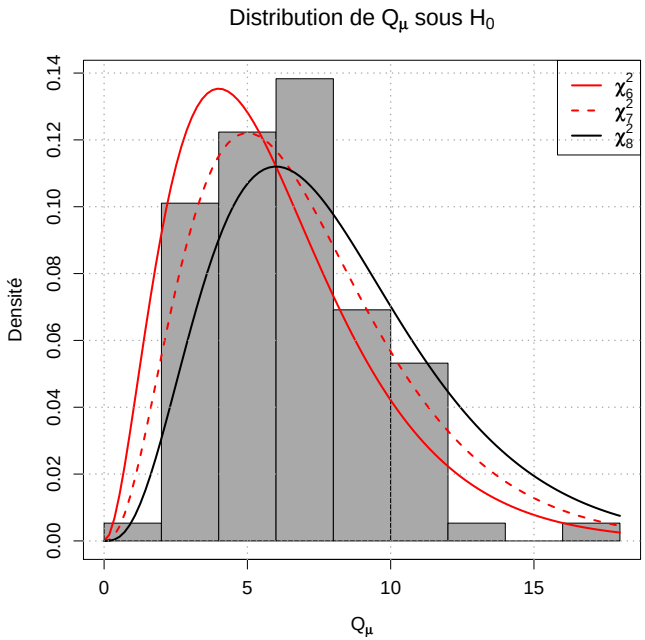
- $\widehat{S}$ matrice diagonale de dimension $J \times J$, dont l'élément (j,j) est $\overline{A_j}$ (la moyenne de $\Delta\widehat{A}_i(t_{i,j})$).
- $\mathrm{H}_0 : \mu = 0$. Sous H_0 $\mathbb{E}\left[\Delta\widehat{M}_i(t_{i,j})\right] \cong 0$.
- Statistique de test :

$$Q_\mu = n\overline{\Delta M}'\widehat{S}^{-1}\overline{\Delta M}$$

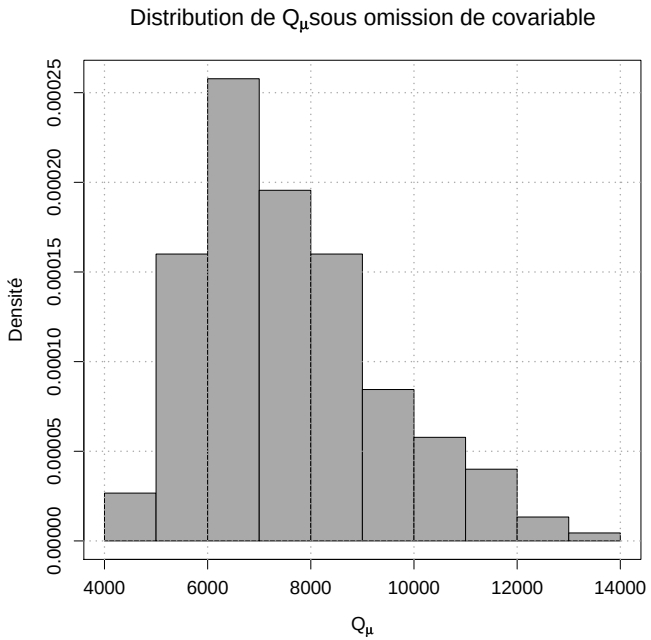
- Sous H_0 asymptotiquement $Q_\mu \sim \chi^2_{(J-1)}$.

GOF et martingales : exemple sur les données simulées

- LEYP avec $\theta = \{1.5,\ 2,\ -11,\ 0.2\}$, 10000 individus, 30 ans d'observation, 100 jeux de données.



Modèle bien spécifié.



Modèle sans covariable.

Nombre maximal de défaillances (pourcentage de jeux de données concernés).

	2	3	4
Jeux de données	38%	58%	4%

Plan

Conclusion et perspectives

- LEYP : cf. Récapitulatif (p. 21).
- Comparaison avec d'autres modèles :
  - Difficile de comparer avec les modèles d'âge virtuel (autres modèles (ARI, ARA$_m$, ...) ?).
  - Comparaison des modèles sur l'impact de la réparation avec les modèles de fragilité ?
- Tests d'ajustement.
  - Présence d'une covariable ?
  - Forme de l'intensité initiale $\lambda_0(.)$?
  - ...

Bibliography I

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993).
Statistical models based on counting process.
Springer-Verlag.

Corset, F., Despréaux, S., Doyen, L., and Gaudoin, O. (2009).
MARS : a software tool for Maintenance Assessment of Repairable Systems.
In *Mathematical methods in reliability. Theory. Methods. Applications*, Moscow.

Doyen, L. and Gaudoin, O. (2004).
Classes of imperfect repair models based on reduction of failure intensity or virtual age.
Reliability Engineering & System Safety, 84(1) :45–56.

Jones, C. and Harrington, D. (2001).
Omnibus Tests of The Martingale Assumption in The Analysis of Recurrent Failure Time Data.
Lifetime Data Analysis, 7(2) :157–171.

Bibliography II

Le Gat, Y. (2009).
Une extension du processus de Yule pour la modélisation stochastique des événements récurrents.
PhD thesis, ENGREF.

Peña, E. (2006).
Dynamic modelling and statistical analysis of event times.
Statistical science : a review journal of the Institute of Mathematical Statistics, 21(4) :1.

Ross, S. (1983).
Stochastic Processes.
John Wiley and Sons, New York.

Stocker, R. and Peña, E. (2007).
A general class of parametric models for recurrent event data.
Technometrics, 49(2) :210–221.