

Risque et assurance

Arthur Charpentier

Université Rennes 1

arthur.charpentier@univ-rennes1.fr

[http ://freakonometrics.blog.free.fr/](http://freakonometrics.blog.free.fr/)



Journées d'Études Statistique, Luminy, Novembre 2010.

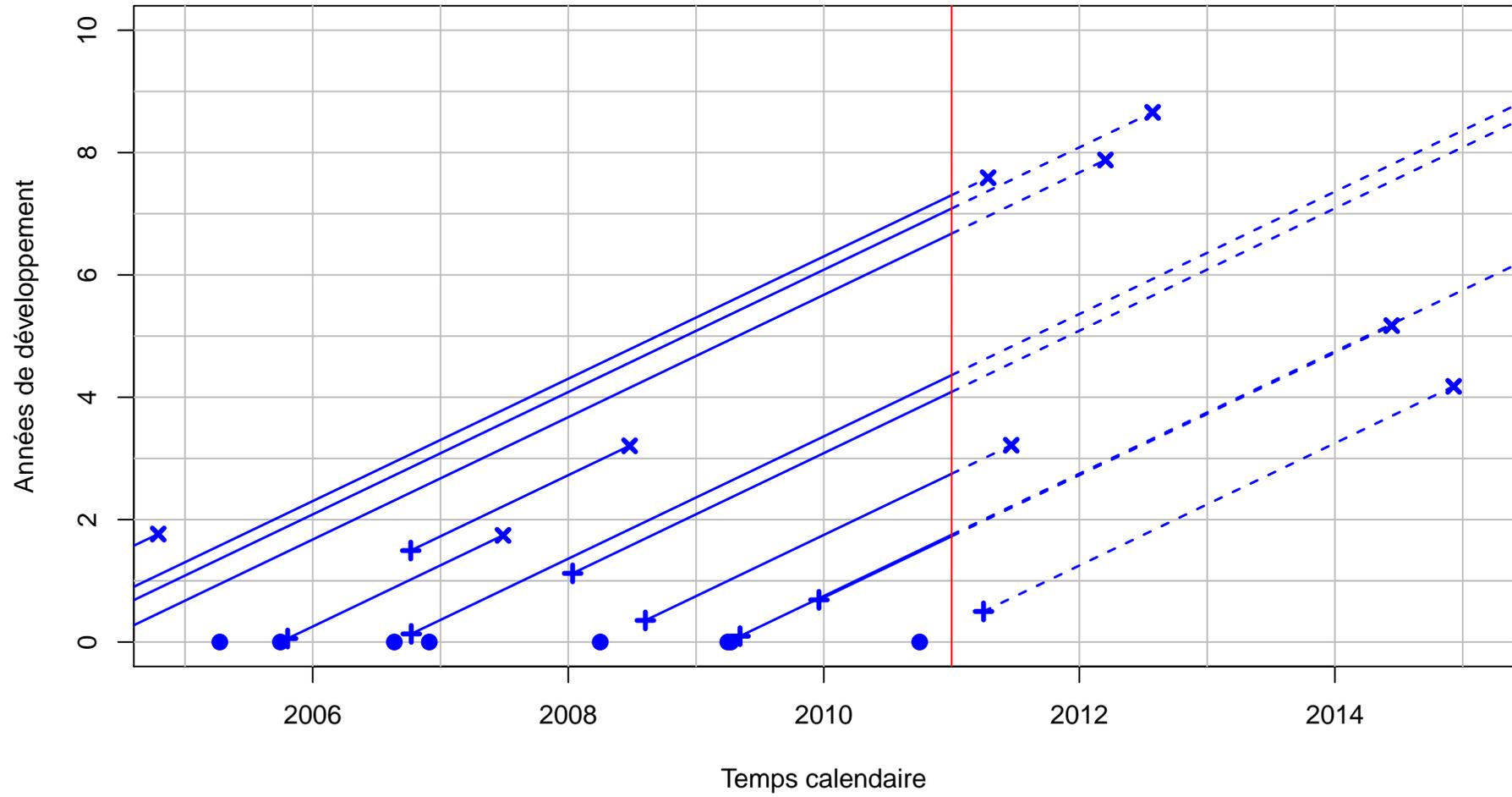
1 Prise en compte de la dynamique en assurance

1.1 La problématique du provisionnement en assurance

[25], “*les provisions techniques sont les provisions destinées à permettre le règlement intégral des engagements pris envers les assurés et bénéficiaires de contrats. Elles sont liées à la technique même de l’assurance, et imposées par la réglementation*”.

A la date t , la compagnie d’assurance est tenue de constituer une **provision** pour les sinistres survenus avant la date t qu’elle sera tenu d’indemniser.

cf. diagramme de Lexis de la vie des sinistres,



| Réglements en | n | $n + 1$ | $n + 2$ | $n + 3$ | $n + 4$ |
|------------------------|-----|---------|---------|---------|---------|
| Multirisque habitation | 55% | 90% | 94% | 95% | 96% |
| Automobile | 55% | 79% | 84% | 99% | 90% |
| <i>dont corporel</i> | 13% | 38% | 50% | 65% | 72% |
| Responsabilité civile | 10% | 25% | 35% | 40% | 45% |

TABLE 1 – Cadences de paiements, pour quelques branches d'activité (source [7]).

Swiss Re

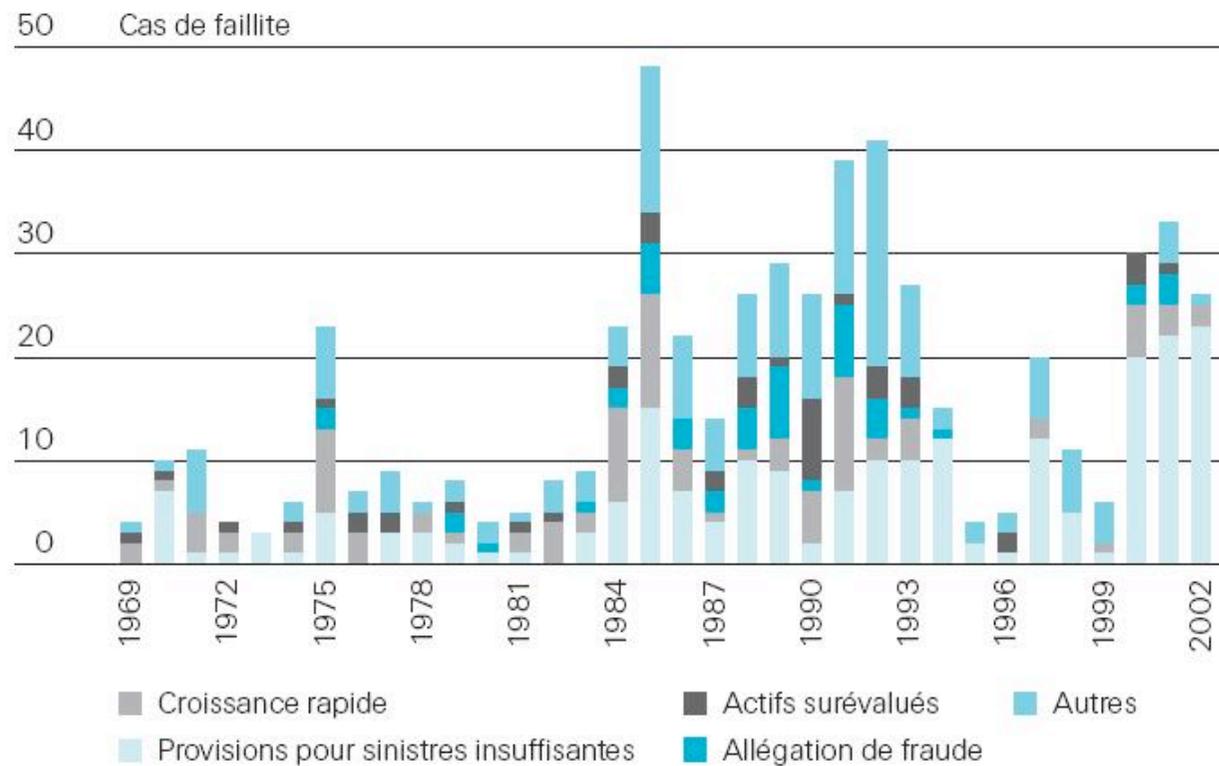


sigma

N° 2/2008

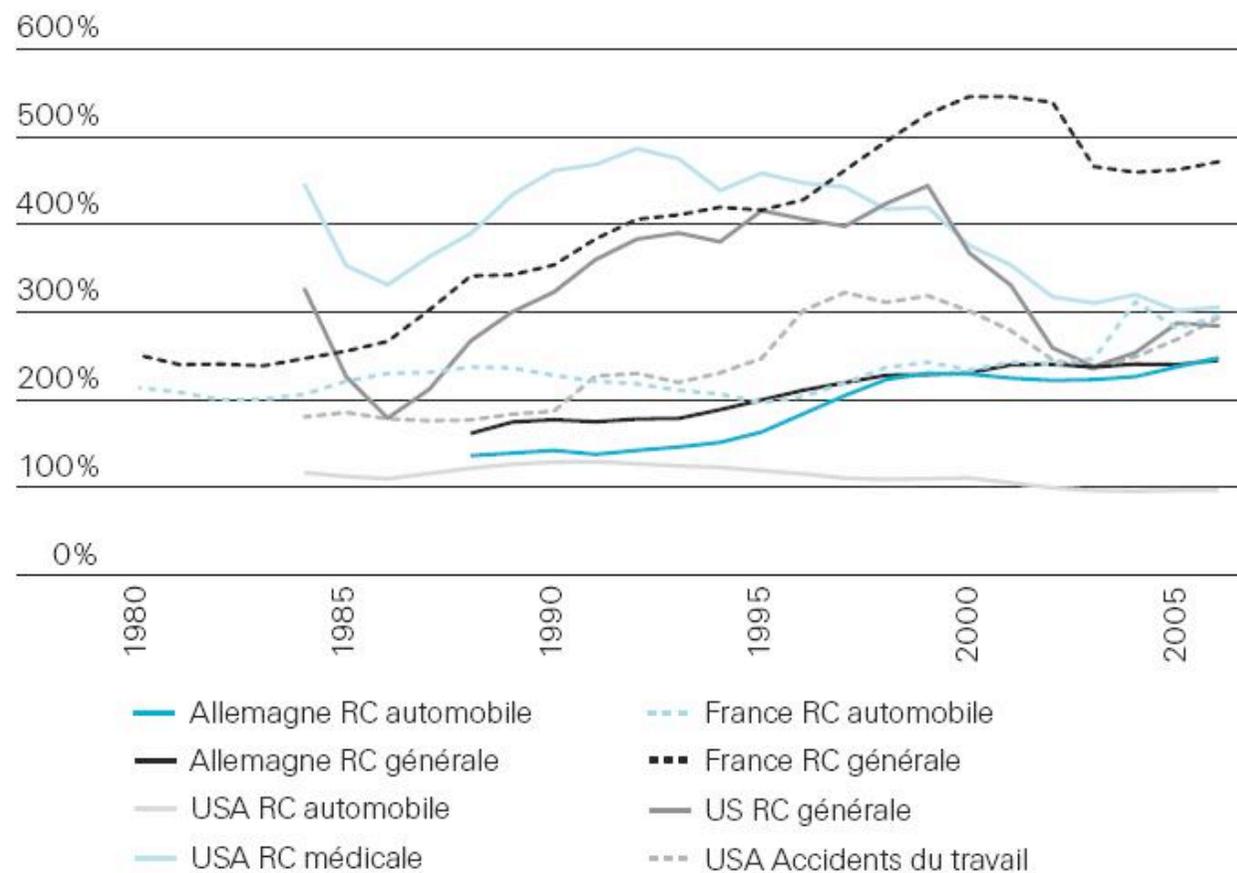
**Provisionnement des sinistres
en assurance de dommages :**
progresser sur un défi stratégique

Graphique 1
 Cas de faillite enregistrés par les assureurs dommages, RC et accidents américains, par cause



Source : A.M. Best : Best's Insolvency Study, Property/Casualty U.S. Insurers 1969-2002, mai 2004, p. 34

Graphique 3
Provisions pour sinistres, en % des primes acquises, 1980-2006



Sources : Autorités de surveillance des assurances, Best's Aggregates & Averages, Property/Casualty, édition 2007

1.2 Quelques définitions et notations, aspects règlementaires et comptables

On notera

- i (en ligne) l'année de survenance, $i = 1, \dots, n$,
- j (en colonne) l'année de développement, $j = 0, \dots, n - 1$,
- $Y_{i,j}$ les **incréments de paiements**, pour l'année de développement j , pour les sinistres survenus l'année i , tableau 2
- $C_{i,j}$ les **paiements cumulés**, au sens où $C_{i,j} = Y_{i,0} + Y_{i,1} + \dots + Y_{i,j}$, pour l'année de survenance j , tableau 3
- P_i la **prime** acquise pour l'année i , tableau 4
- $N_{i,j}$ le **nombre cumulé de sinistres** pour l'année de survenance i vu au bout de j années, tableau 5

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|------|------|-----|----|----|----|
| 1 | 3209 | 1163 | 39 | 17 | 7 | 21 |
| 2 | 3367 | 1292 | 37 | 24 | 10 | |
| 3 | 3871 | 1474 | 53 | 22 | | |
| 4 | 4239 | 1678 | 103 | | | |
| 5 | 4929 | 1865 | | | | |
| 6 | 5217 | | | | | |

TABLE 2 – Triangle des incréments de paiements, $\mathbf{Y} = (Y_{i,j})$.

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|------|------|------|------|------|------|
| 1 | 3209 | 4372 | 4411 | 4428 | 4435 | 4456 |
| 2 | 3367 | 4659 | 4696 | 4720 | 4730 | |
| 3 | 3871 | 5345 | 5398 | 5420 | | |
| 4 | 4239 | 5917 | 6020 | | | |
| 5 | 4929 | 6794 | | | | |
| 6 | 5217 | | | | | |

TABLE 3 – Triangle des paiements cumulés, $\mathbf{C} = (C_{i,j})$.

| Year i | 1 | 2 | 3 | 4 | 5 | 6 |
|----------|------|------|------|------|------|------|
| P_i | 4591 | 4672 | 4863 | 5175 | 5673 | 6431 |

TABLE 4 – Vecteur des primes acquises, $\mathbf{P} = (P_i)$.

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|--------|--------|--------|--------|--------|--------|
| 1 | 1043.4 | 1045.5 | 1047.5 | 1047.7 | 1047.7 | 1047.7 |
| 2 | 1043.0 | 1027.1 | 1028.7 | 1028.9 | 1028.7 | |
| 3 | 965.1 | 967.9 | 967.8 | 970.1 | | |
| 4 | 977.0 | 984.7 | 986.8 | | | |
| 5 | 1099.0 | 1118.5 | | | | |
| 6 | 1076.3 | | | | | |

TABLE 5 – Triangle des nombres de sinistres, cumulés, en milliers, $\mathbf{N} = (N_{i,j})$.

1.3 Formalisation du problème du provisionnement

Le provisionnement est un problème de **prédiction**, conditionnelle à \mathcal{H}_n l'information disponible à la date n ,

$$\mathcal{H}_n = \{(Y_{i,j}), i + j \leq n\} = \{(C_{i,j}), i + j \leq n\}$$

Si l'on se focalise sur une année de survenance particulière, on pourra noter

$$\mathcal{F}_{i,n-i} = \{(Y_{i,j}), j = 0, \dots, n - i\} = \{(C_{i,j}), j = 0, \dots, n - i\}$$

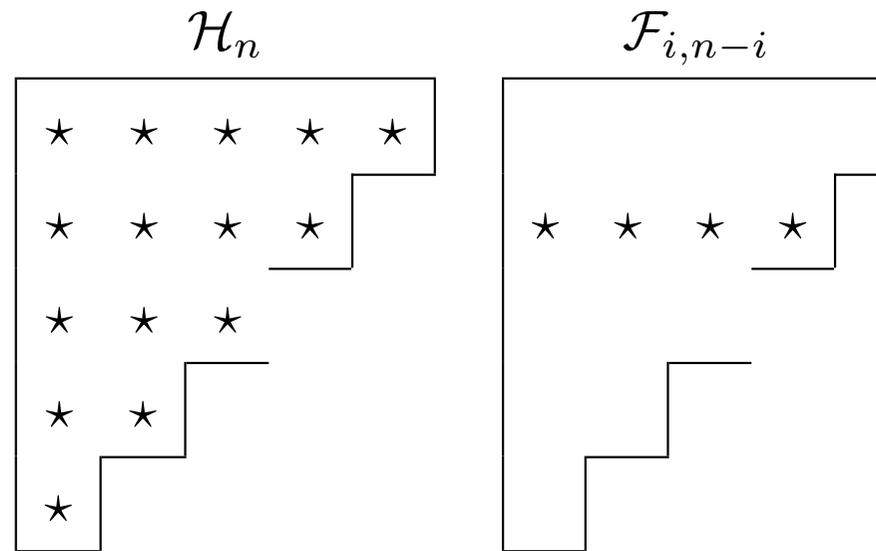


FIGURE 1 – Les informations disponibles pour faire de la prédiction.

On cherchera par la suite à prédire le montant des sinistres à payer pour l'année i , i.e.

$$\widehat{C}_{i,n}^{(n-i)} = \mathbb{E}[C_{i,n} | \mathcal{F}_{i,n-i}]$$

et la différence entre ce montant et le montant déjà payé constituera la provision pour sinistres à payer,

$$\widehat{R}_i = \widehat{C}_{i,n}^{(n-i)} - C_{i,n-i}.$$

On cherche

$$\text{mse}[C_{i,n} | \mathcal{F}_{i,n-i}] \text{ ou } \text{mse}[\widehat{C}_{i,n}^{(n-i)}],$$

ce que l'on appellera **incertitude à horizon ultime**.

Solvabilité II : mesurer une **incertitude à un an**. Dans un an, on prédira

$$\widehat{C}_{i,n}^{(n-i+1)} = \mathbb{E}[C_{i,n} | \mathcal{F}_{i,n-i+1}]$$

i.e. changement dans l'estimation de la charge ultime

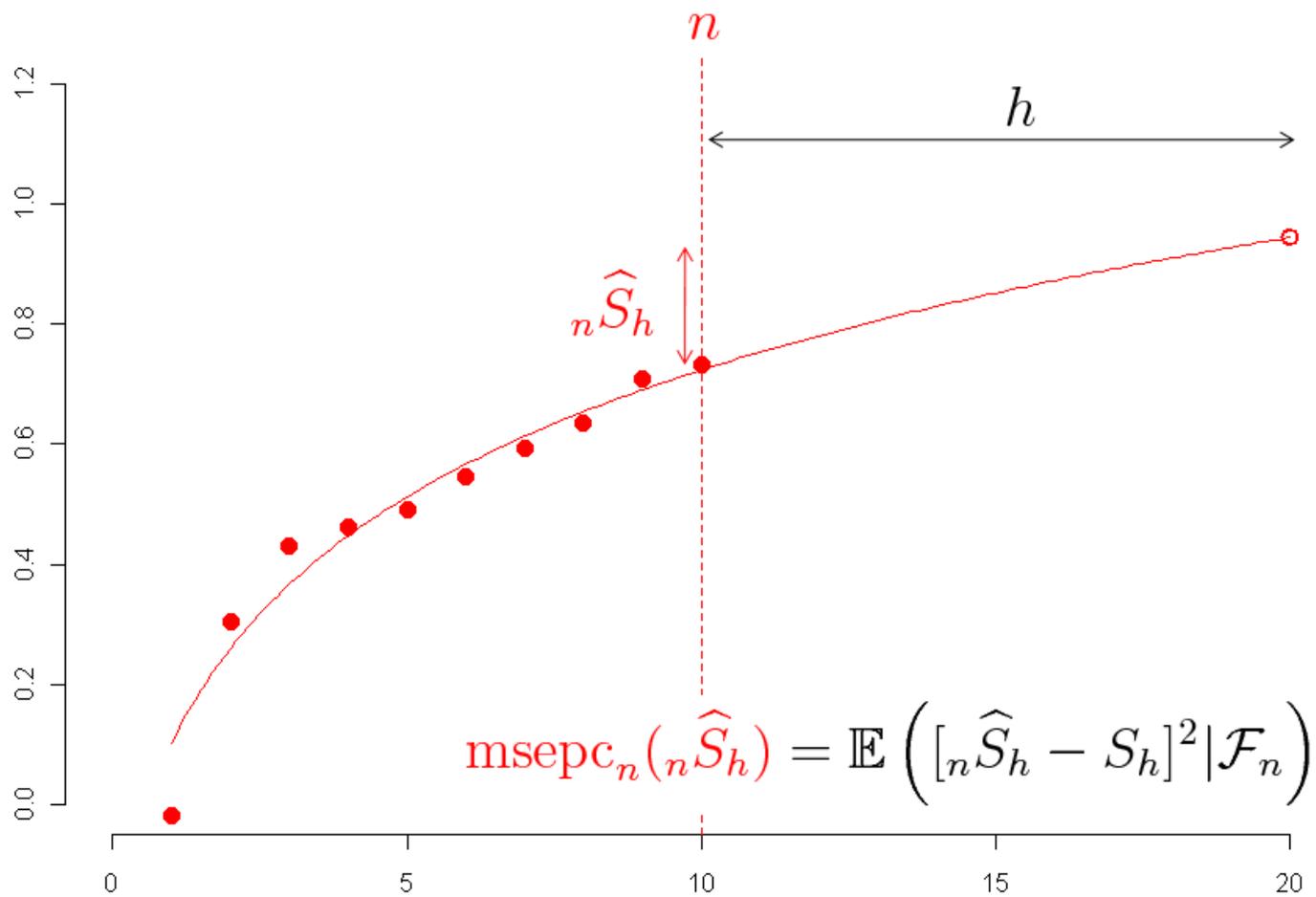
$$\Delta_i^n = \widehat{C}_{i,n}^{(n-i+1)} - \widehat{C}_{i,n}^{(n-i)}.$$

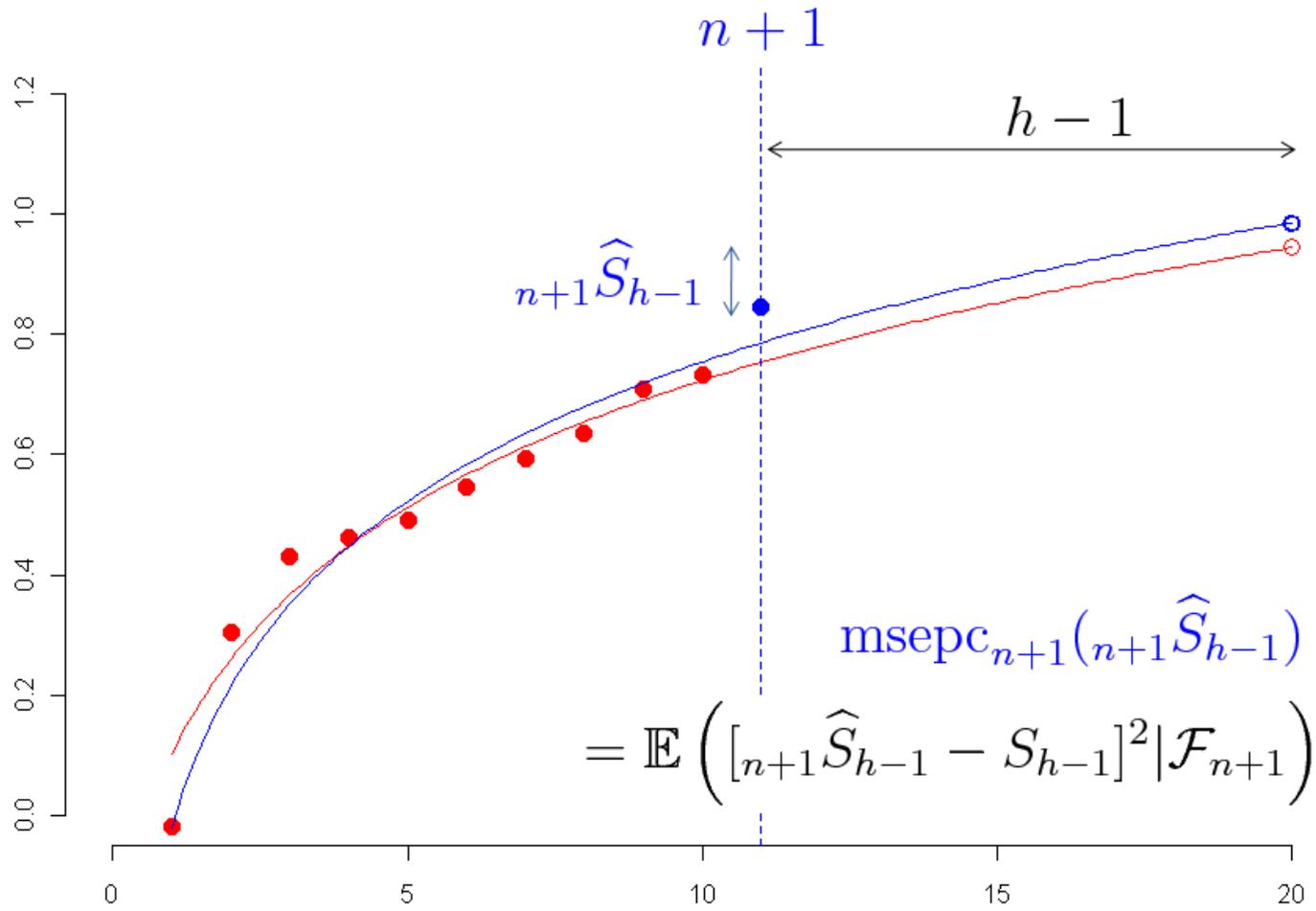
Si cette différence est positive, on parle de **mali** (il faudra gonfler la provision afin de pouvoir payer les sinistres), et si elle est négative, on parle de **boni**.

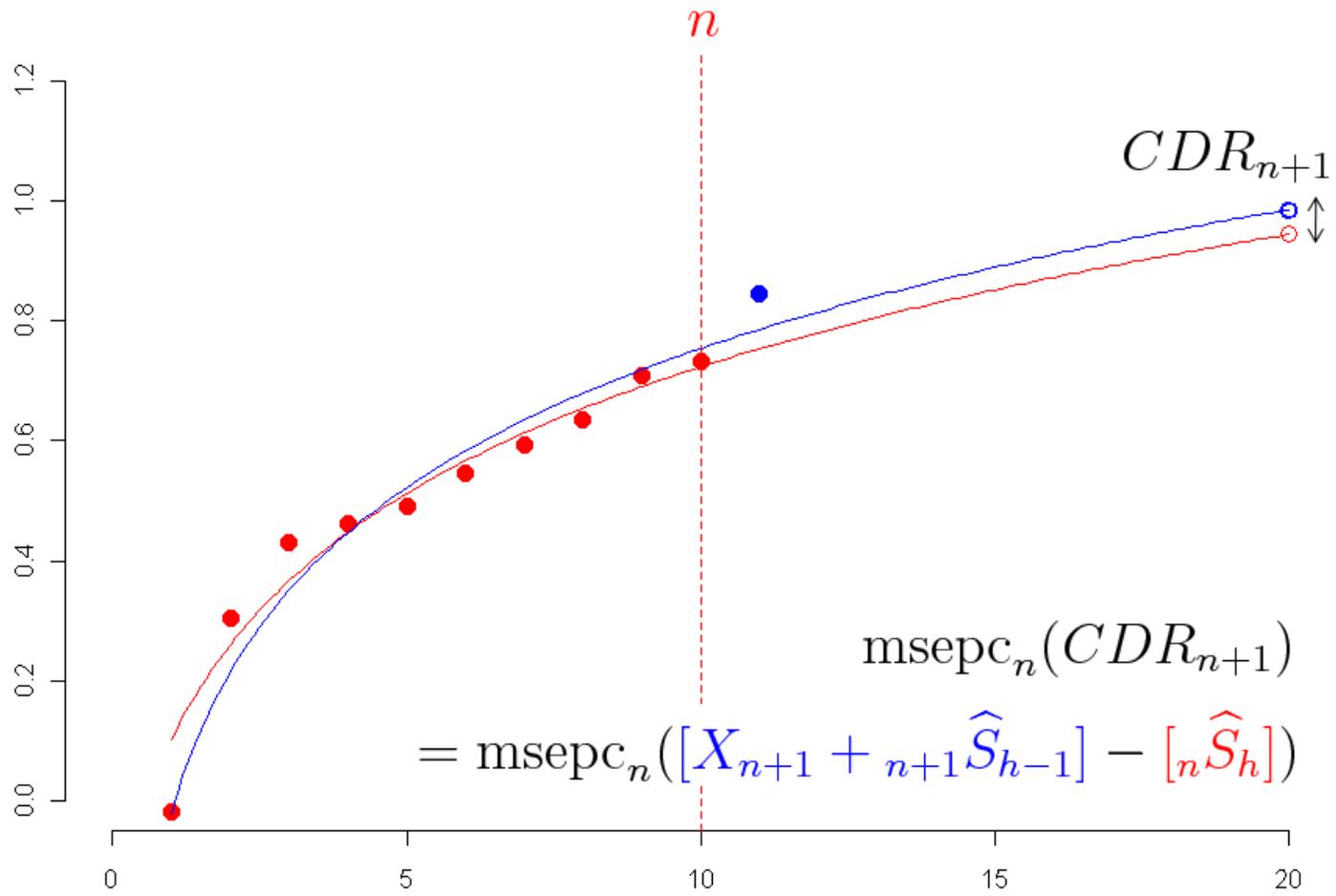
$$\mathbb{E}[\Delta_i^n | \mathcal{F}_{i,n-i}] = 0,$$

Solvabilité II : calculer

$$\text{mse}[\Delta_i^n | \mathcal{F}_{i,n-i}].$$







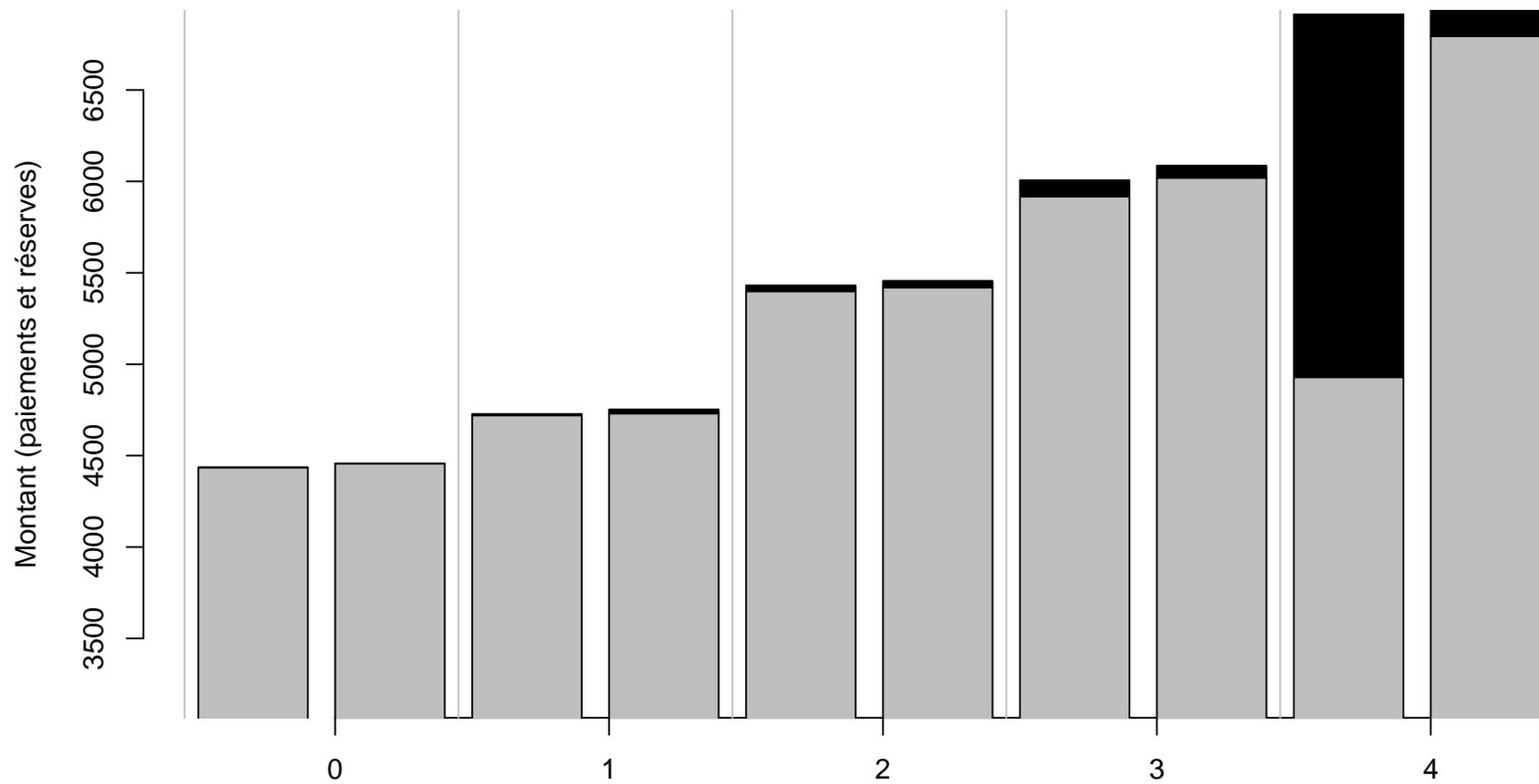


FIGURE 2 – Estimation de la charge ultime $\hat{C}_{i,n}$ deux années consécutives.

1.4 Lecture transversale et longitudinale des tables de mortalité

démographie & assurance-vie, probabilité qu'une personne en vie à l'âge x soit en vie à l'âge $x + h$

$$\mathbb{P}(T > x + h | T > x) = \frac{L_{x+h}}{L_x}$$

où L_j désigne le **nombre de survivants** d'âge j , dans une table de mortalité.

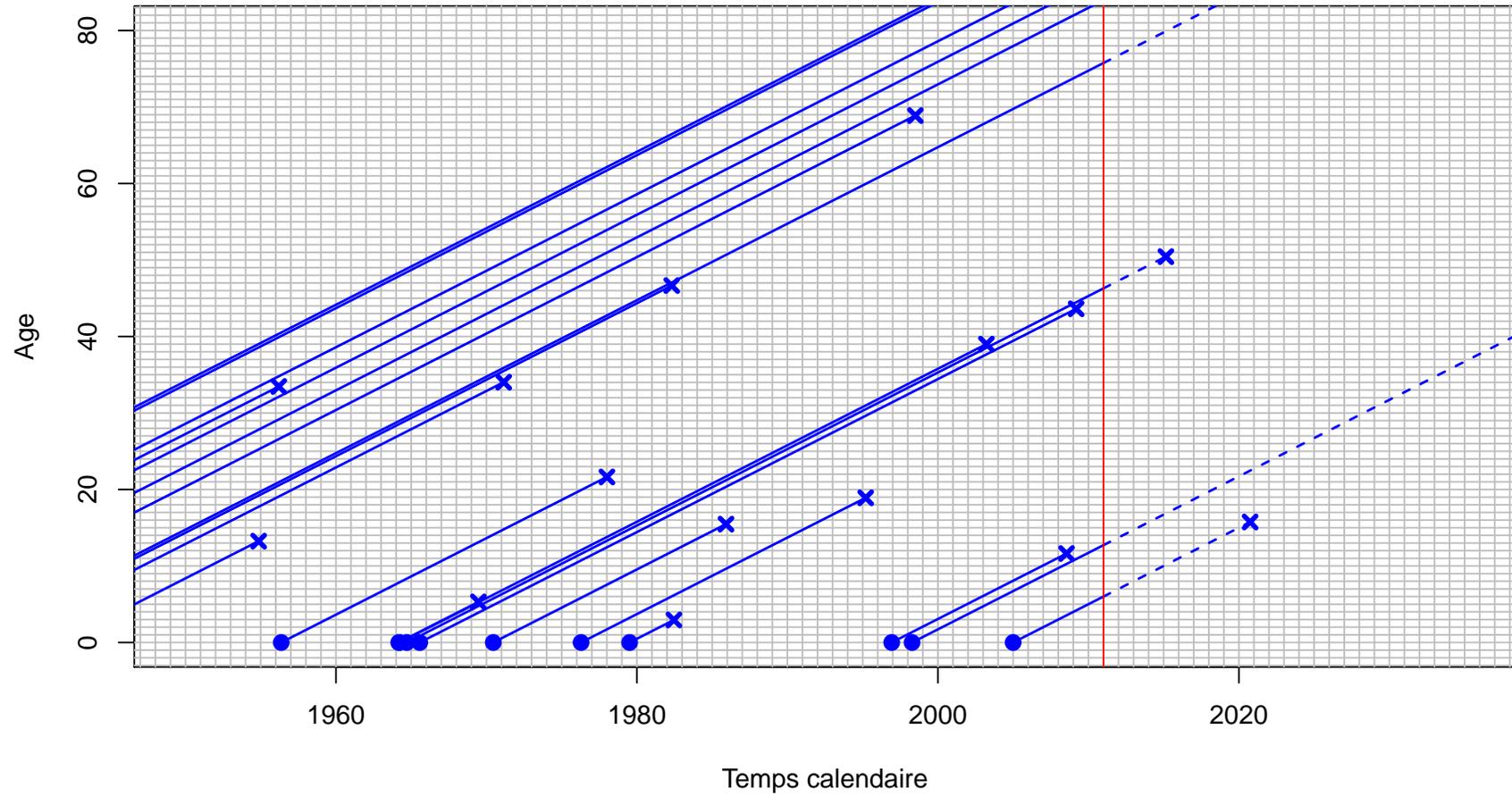
Si on se place l'année t , la probabilité précédente *devrait* s'écrire

$$\begin{aligned} \mathbb{P}_t(T > x + h | T > x) = & \mathbb{P}_t(T > x + 1 | T > x) \cdot \mathbb{P}_{t+1}(T > x + 2 | T > x + 1) \cdots \\ & \cdots \mathbb{P}_{t+h-1}(T > x + h | T > x + h - 1) \end{aligned}$$

soit, si $L_{t,j}$ désigne le nombre de survivants d'âge j , observés l'année t ,

$$\mathbb{P}_t(T > x + h | T > x) = \frac{L_{t,x+1}}{L_{t,x}} \cdot \frac{L_{t+1,x+2}}{L_{t+1,x+1}} \cdots \frac{L_{t+h-1,x+h}}{L_{t+h-1,x+h-1}}.$$

Besoin d'**extrapoler** ces données.

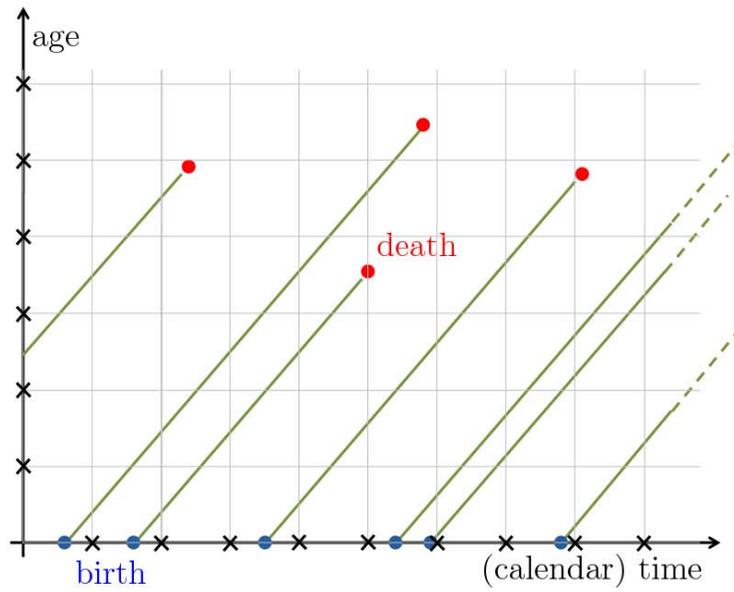


Modèle de **Lee & Carter**, [15] : on suppose pour cela que le **taux de décès**, à la date t pour une personne d'âge x dépend de trois facteurs,

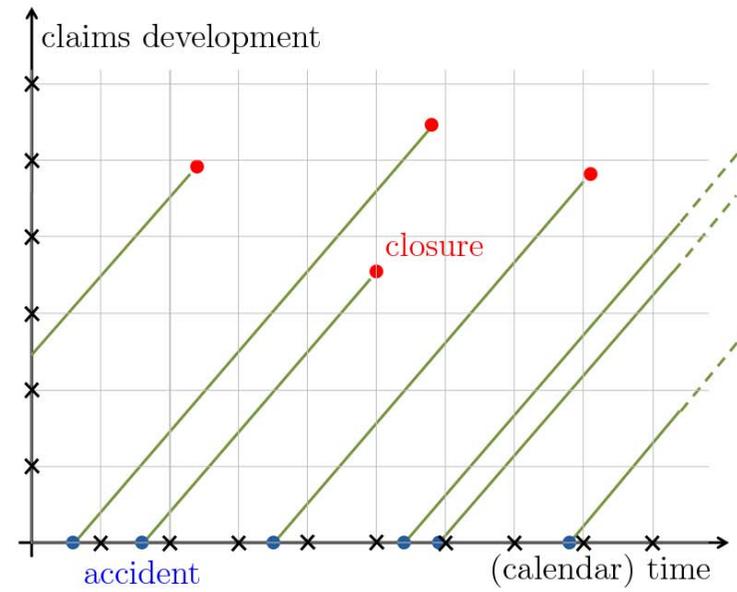
$$\mu_{x,t} = \exp[\alpha_x + \beta_x \cdot \kappa_t],$$

où $\alpha = (\alpha_x)$ désigne l'évolution *moyenne* de la mortalité en fonction de l'âge, et où $\kappa = (\kappa_t)$ permet de modéliser l'évolution de la mortalité dans le temps (traduisant une globale amélioration des conditions de vie). Toutefois, en multipliant par $\beta = (\beta_x)$, on peut prendre en compte le fait que les gains ne sont pas identiques pour tous les âges.

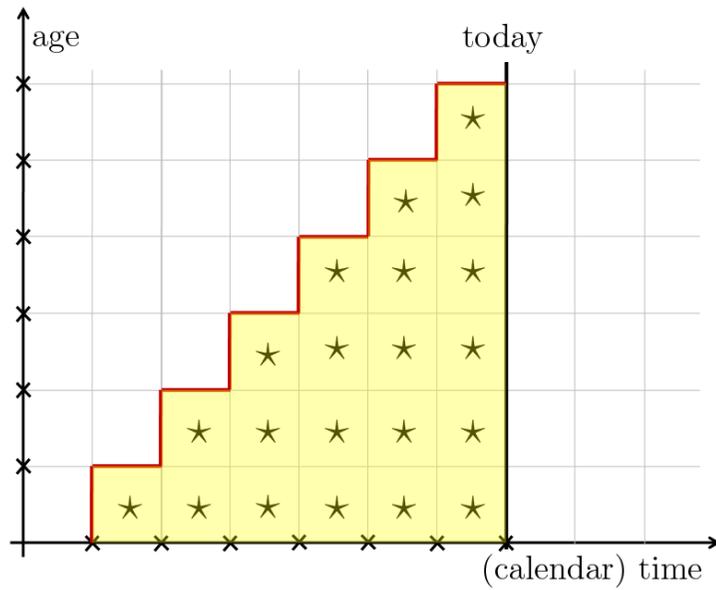
in **life** insurance



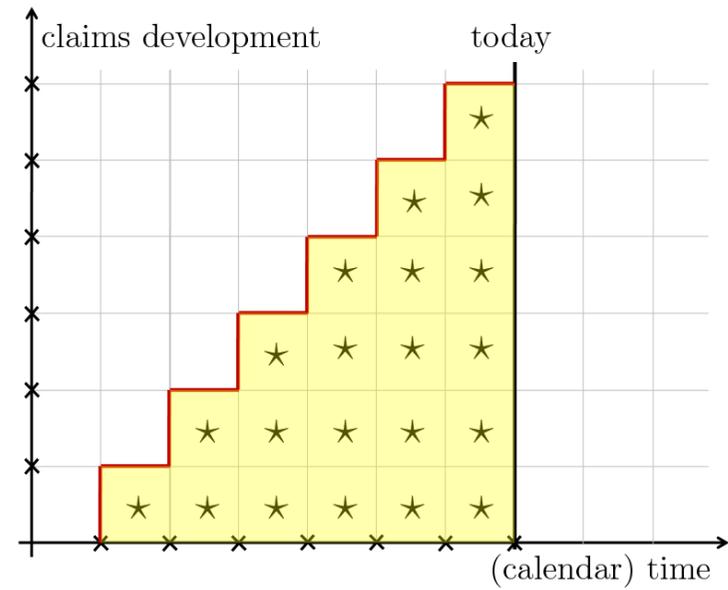
in **nonlife** insurance



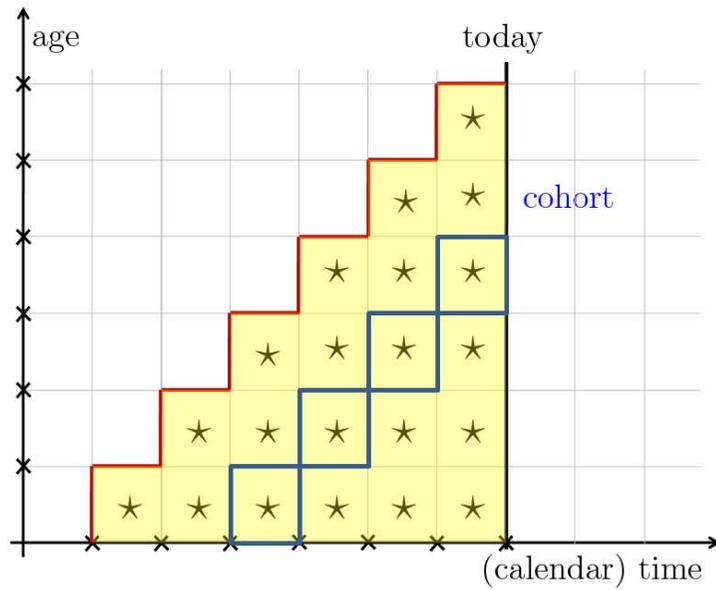
in **life** insurance



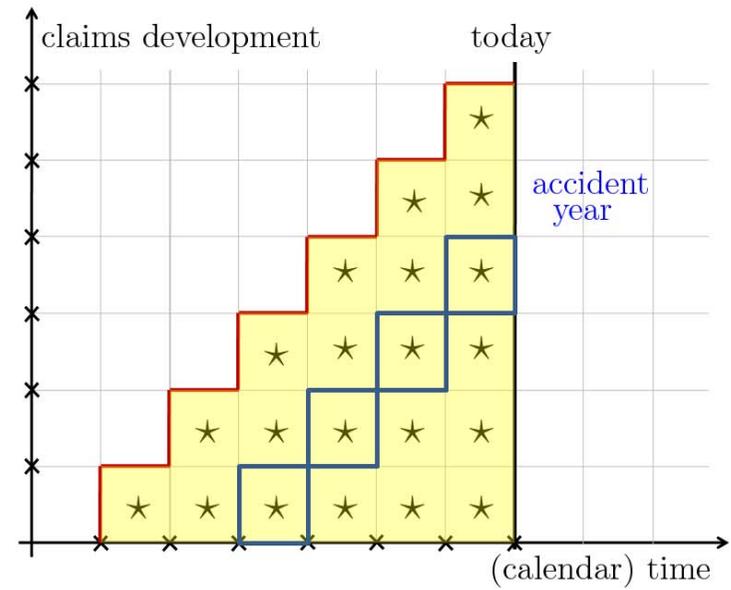
in **nonlife** insurance



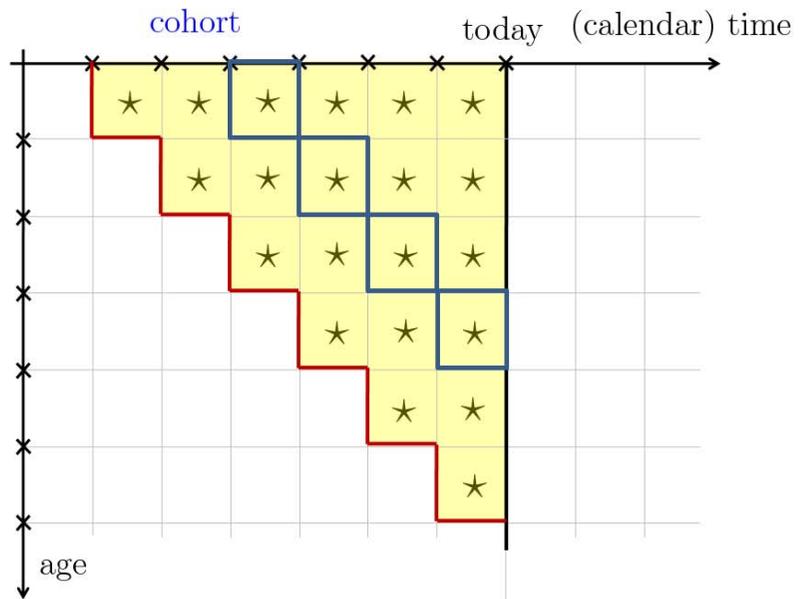
in **life** insurance



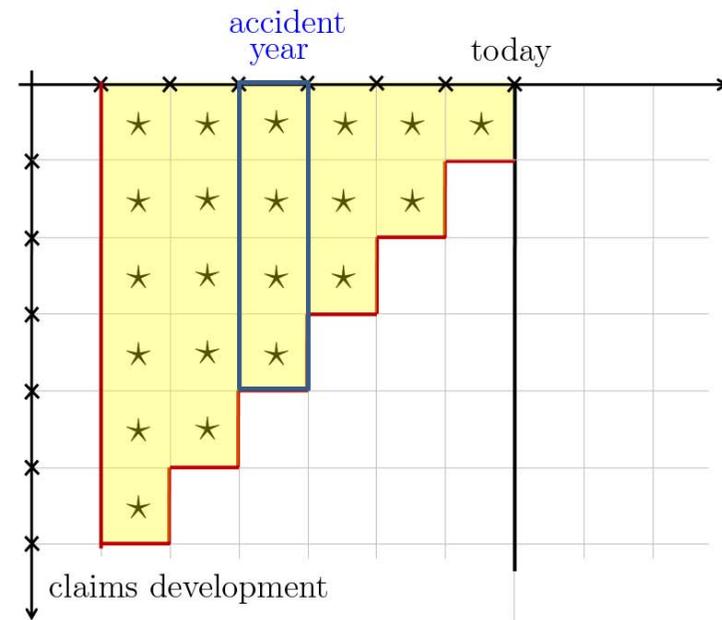
in **nonlife** insurance



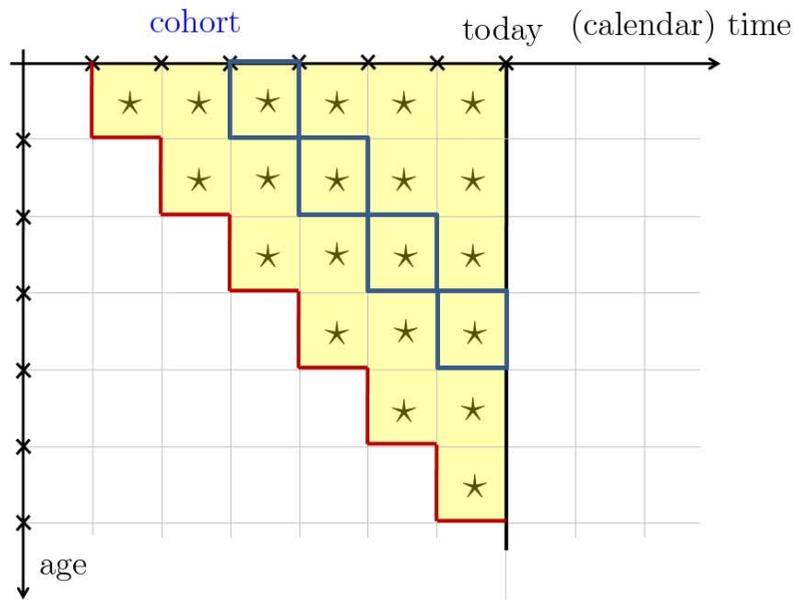
in **life** insurance



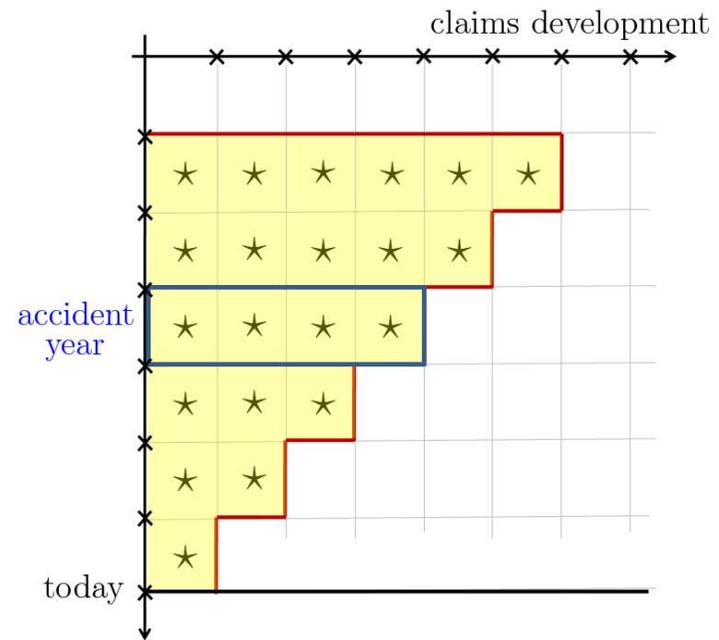
in **nonlife** insurance



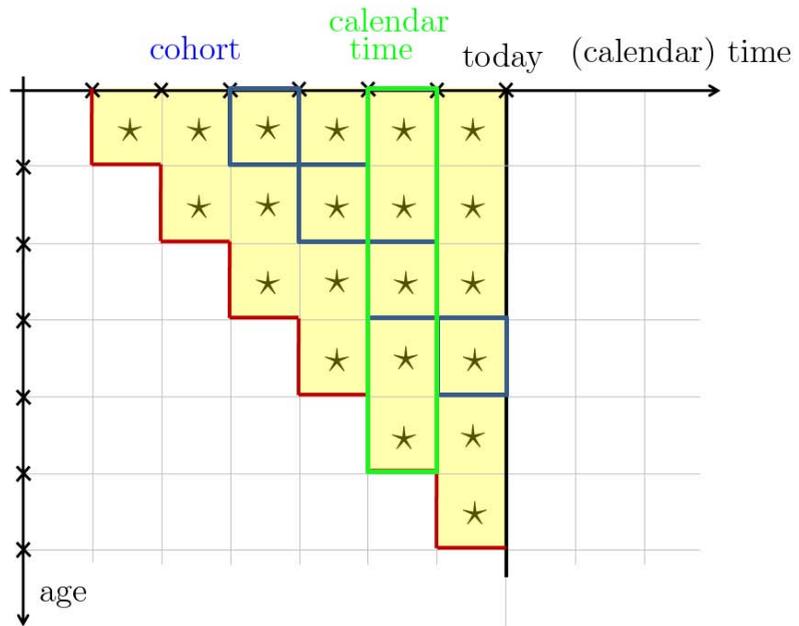
in **life** insurance



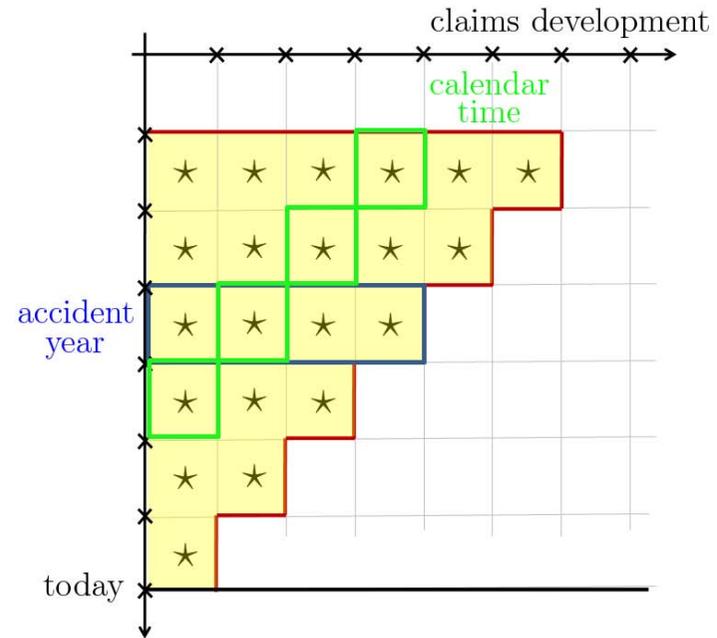
in **nonlife** insurance



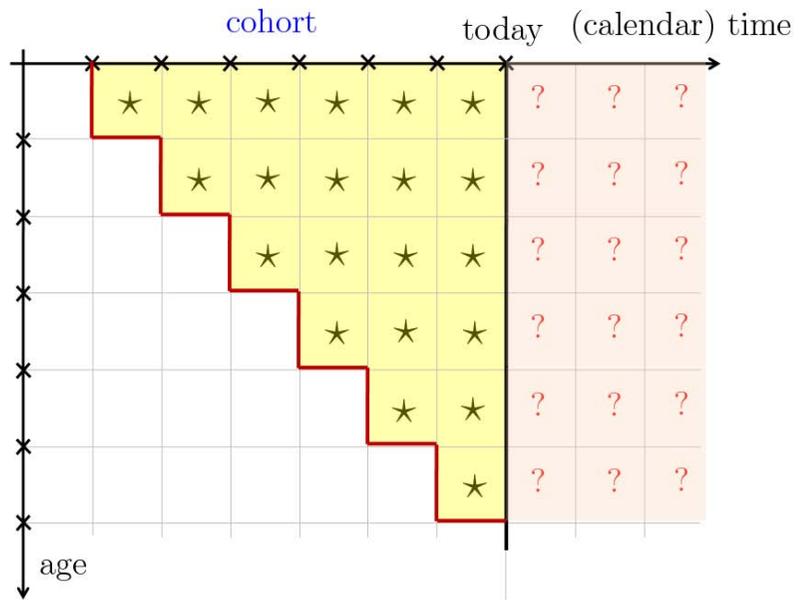
in **life** insurance



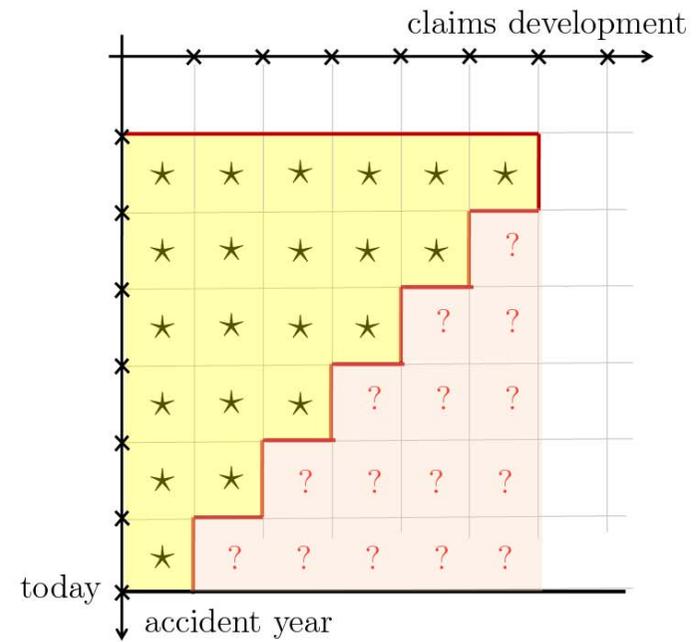
in **nonlife** insurance



in **life** insurance



in **nonlife** insurance



2 Les cadences de paiements et la méthode Chain Ladder

L'utilisation des cadences de paiements pour estimer la charge future date des années 30. On suppose qu'il existe une relation de récurrence de la forme

$$C_{i,j+1} = \lambda_j \cdot C_{i,j} \text{ pour tout } i, j = 1, \dots, n.$$

Un estimateur naturel pour λ_j , basé sur l'expérience passée est alors

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{i,j}} \text{ pour tout } j = 1, \dots, n-1.$$

Il s'agit tout simplement du ratio moyen entre les années de développement j et $j+1$. De telle sorte que l'on peut alors prédire la charge pour la partie non-observée dans le triangle,

$$\hat{C}_{i,j} = \left[\hat{\lambda}_{n-i} \dots \hat{\lambda}_{j-1} \right] C_{i,n+1-i}.$$

2.1 Des réécritures du modèle

Notons qu'au lieu de calculer les facteurs de développement, on peut aussi des taux de développement, cumulés ou non. Autrement dit, au lieu d'écrire $C_{i,j+1} = \lambda_j C_{i,j}$ pour tout $i, j = 1, \dots, n$, on suppose que

$$C_{i,j} = \gamma_j C_{i,n} \text{ ou } Y_{i,j} = \varphi_j C_{i,n}.$$

On notera que

$$\gamma_j = \prod_{k=j+1}^n \frac{1}{\lambda_k} \text{ et } \varphi_j = \begin{cases} \gamma_1 & \text{si } j = 1 \\ \gamma_j - \gamma_{j-1} & \text{si } j > 1 \end{cases}$$

Ces valeurs ont été calculées dans le Tableau 6.

| | 0 | 1 | 2 | 3 | 4 | 5 |
|-------------|---------|---------|---------|---------|---------|----------|
| λ_j | 1,38093 | 1,01143 | 1,00434 | 1,00186 | 1,00474 | 1,0000 |
| γ_j | 70,819% | 97,796% | 98,914% | 99,344% | 99,529% | 100,000% |
| φ_j | 70,819% | 26,977% | 1,118% | 0,430% | 0,185% | 0,000% |

TABLE 6 – Facteurs de développement, $\hat{\lambda} = (\hat{\lambda}_i)$, exprimés en cadence de paiements par rapport à la charge ultime, en cumulé (i.e. $\hat{\gamma}$), puis en incréments (i.e. $\hat{\varphi}$).

Enfin, un peu de réécriture montre qu'il est possible de voir l'estimateur Chain-Ladder comme une moyenne pondérée des facteurs de transition individuels, i.e.

$$\hat{\lambda}_j = \sum_{i=1}^{n-j} \omega_{i,j} \lambda_{i,j} \quad \text{où} \quad \omega_{i,j} = \frac{C_{i,j}}{\sum_{i=1}^{n-j} C_{i,j}} \quad \text{et} \quad \lambda_{i,j} = \frac{C_{i,j+1}}{C_{i,j}}.$$

Aussi, on peut obtenir ces coefficients à l'aide de régressions linéaires pondérées

sans constantes, en régressant les $C_{\cdot,j+1}$ sur les $C_{\cdot,j}$,

$$\hat{\lambda}_j = \operatorname{argmin}_{\lambda \in \mathbb{R}} \left\{ \sum_{i=1}^{n-j} C_{i,j} \left[\lambda - \frac{C_{i,j+1}}{C_{i,j}} \right]^2 \right\},$$

ou encore

$$\hat{\lambda}_j = \operatorname{argmin}_{\lambda \in \mathbb{R}} \left\{ \sum_{i=1}^{n-j} \frac{1}{C_{i,j}} [\lambda C_{i,j} - C_{i,j+1}]^2 \right\}.$$

A partir du calcul des facteurs de transition $\hat{\lambda} = (\hat{\lambda}_i)$, on complète alors le triangle, en remontant d'année de survenance en année de survenance, en commençant par $j = 1$, puis $j = 2$, etc (d'échelon en échelon, d'où le nom de cette méthode). Le triangle complété est présenté dans le Tableau 2.1.

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|------|--------|--------|---------|--------|--------|
| 1 | 3209 | 4372 | 4411 | 4428 | 4435 | 4456 |
| 2 | 3367 | 4659 | 4696 | 4720 | 4730 | 4752.4 |
| 3 | 3871 | 5345 | 5398 | 5420 | 5430.1 | 5455.8 |
| 4 | 4239 | 5917 | 6020 | 6046.15 | 6057.4 | 6086.1 |
| 5 | 4929 | 6794 | 6871.7 | 6901.5 | 6914.3 | 6947.1 |
| 6 | 5217 | 7204.3 | 7286.7 | 7318.3 | 7331.9 | 7366.7 |

TABLE 7 – Triangle des paiements cumulés, $\mathbf{C} = (C_{i,j})_{i+j \leq n}$ avec leur projection future $\hat{\mathbf{C}} = (\hat{C}_{i,j})_{i+j > n}$.

2.2 Une approche par sommes marginales

Au delà de l'approche Markovienne du problème de provisionnement que l'on vient de voir (et qui sera reprise dans la section 3), il existe une caractérisation particulièrement intéressante, que l'on retrouvera dans la section 4.

En fait, si l'on cherche des vecteurs $\mathbf{A} = (A_0, \dots, A_n)$ et $\mathbf{B} = (B_0, \dots, B_n)$, avec $B_0 + \dots + B_n = 1$, tels que

$$\sum_{i=1}^{n-j} A_i B_j = \sum_{i=1}^{n-j} Y_{i,j} \text{ pour tout } j, \text{ et } \sum_{j=0}^{n-i} A_i B_j = \sum_{j=0}^{n-i} Y_{i,j} \text{ pour tout } i,$$

(on ne somme que sur la partie observée du triangle) les montants prédits dans la partie inférieure du triangles, i.e. $(A_i B_j)_{i+j > n}$, coïncident avec les quantités prédites par la méthode Chain Ladder ([23]).

Proposition 1

S'il existe $\mathbf{A} = (A_0, \dots, A_n)$ et $\mathbf{B} = (B_0, \dots, B_n)$, avec $B_0 + \dots + B_n = 1$, tels que

$$\sum_{i=1}^{n-j} A_i B_j = \sum_{i=1}^{n-j} Y_{i,j} \text{ pour tout } j, \text{ et } \sum_{j=0}^{n-i} A_i B_j = \sum_{j=0}^{n-i} Y_{i,j} \text{ pour tout } i,$$

alors

$$\hat{C}_{i,n} = A_i = C_{i,n-i} \cdot \prod_{k=n-i}^{n-1} \lambda_k$$

où

$$B_k = \prod_{j=k}^{n-1} \frac{1}{\lambda_j} - \prod_{j=k-1}^{n-1} \frac{1}{\lambda_j}, \text{ avec } B_0 = \prod_{j=0}^{n-1} \frac{1}{\lambda_j}.$$

Autrement dit, le montant de provision coïncide avec l'estimateur obtenu par la méthode Chain Ladder.

cf. méthode des marges ([2]).

Remarque1

L'idée de cette méthode remonte aux années 30. Comme le note [1], décrivant la méthode Chain-Ladder, *"son apparente rigueur mathématique et sa simplicité semblent résoudre sans effort toutes les difficultés. Il n'en est malheureusement rien. [...] La cadence de règlements ne peut être constante que si l'assureur prend en charge un nombre de sinistres suffisant pour que la loi des grands nombres puisse jouer. Les changements de jurisprudence qui aggravent la responsabilité des assurés et les dépréciations monétaires aboutissent à une augmentation des indemnités allouées, et ralentissent la cadence des règlements des exercices antérieurs dont les sinistres non encore réglés se trouvent majorés". Et plus précisément, [11] remarquait que "la méthode de la cadence numérique appliquée aux éléments statistiques que nous possédons nous paraît donc devoir donner des résultats empreints d'une assez grande incertitude."*

3 De Mack à Merz & Wüthrich

3.1 Quantifier l'incertitude dans une prédiction

On souhaite comparer \hat{R} et R (où R n'est pas inconnu mais aléatoire). Le **mse de prédiction**, peut s'écrire (en faisant une approximation au premier ordre),

$$\mathbb{E}([\hat{R} - R]^2) \approx \underbrace{\mathbb{E}([\hat{R} - \mathbb{E}(R)]^2)}_{\text{mse}(\hat{R})} + \underbrace{\mathbb{E}([R - \mathbb{E}(R)]^2)}_{\text{Var}(R)}$$

En fait, en toute rigueur (et nous en aurons besoin par la suite), on cherche un msep conditionnel à l'information dont on dispose au bout de n années,

$$\text{mse}_{p_n}(\hat{R}) = \mathbb{E}([\hat{R} - R]^2 | \mathcal{H}_n).$$

3.2 Le formalisme de Mack

[17] a proposé un cadre probabiliste afin de justifier l'utilisation de la méthode Chain-Ladder. Pour cela, on suppose que $(C_{i,j})_{j \geq 0}$ est un processus Markovien, et qu'il existe $\lambda = (\lambda_j)$ et $\sigma = (\sigma_j^2)$ tels que

$$\begin{cases} \mathbb{E}(C_{i,j+1} | \mathcal{H}_{i+j}) = \mathbb{E}(C_{i,j+1} | C_{i,j}) = \lambda_j \cdot C_{i,j} \\ \text{Var}(C_{i,j+1} | \mathcal{H}_{i+j}) = \text{Var}(C_{i,j+1} | C_{i,j}) = \sigma_j^2 \cdot C_{i,j} \end{cases}$$

On note que sous ces hypothèses,

$$\mathbb{E}(C_{i,j+k} | \mathcal{H}_{i+j}) = \mathbb{E}(C_{i,j+k} | C_{i,j}) = \lambda_j \cdot \lambda_{j+1} \cdots \lambda_{j+k-1} C_{i,j}$$

[17] rajoute une hypothèse supplémentaire d'**indépendance entre les années de survenance**, $(C_{i,j})_{j=1,\dots,n}$ et $(C_{i',j})_{j=1,\dots,n}$ sont indépendants pour tout $i \neq i'$.

Une réécriture du modèle est alors de supposer que

$$C_{i,j+1} = \lambda_j C_{i,j} + \sigma_j \sqrt{C_{i,j}} \varepsilon_{i,j}$$

où les résidus $(\varepsilon_{i,j})$ sont i.i.d., centrés et de variance unitaire.

⇒ méthodes des moindres carrés pondérés pour estimer ces coefficients, avec des poids inversement proportionnels à la variance, $C_{i,j}$, i.e. à j donné, on cherche à résoudre

$$\min \left\{ \sum_{i=1}^{n-j} \frac{1}{C_{i,j}} (C_{i,j+1} - \lambda_j C_{i,j})^2 \right\}$$

Pour tester ces deux premières hypothèses, on commence par représenter les $C_{.,j+1}$ en fonction des $C_{.,j}$ à j donné. Si la première hypothèse est vérifiée, les points doivent être alignés suivant une droite passant par l'origine.

Pour la seconde hypothèse, on peut étudier les résidus standardisés ([17] parle de *weighted residuals*), $\hat{\varepsilon}_{i,j} = \frac{C_{i,j+1} - \hat{\lambda}_j C_{i,j}}{\sqrt{C_{i,j}}}$.

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{i,j}}, \forall j = 1, \dots, n-1.$$

est un estimateur sans biais de λ_j .

L'utilisation des résidus standardisés nous donnent d'ailleurs une idée simple pour estimer le paramètre de volatilité.

$$\hat{\sigma}_j^2 = \frac{1}{n-j-1} \sum_{i=1}^{n-j-1} \left(\frac{C_{i,j+1} - \hat{\lambda}_j C_{i,j}}{\sqrt{C_{i,j}}} \right)^2$$

ce qui peut aussi s'écrire

$$\hat{\sigma}_j^2 = \frac{1}{n-j-1} \sum_{i=1}^{n-j-1} \left(\frac{C_{i,j+1}}{C_{i,j}} - \hat{\lambda}_j \right)^2 \cdot C_{i,j}$$

(ce qui est à rapprocher de l'écriture du facteur de transition λ comme moyenne pondérée des facteurs de transitions observés).

Cette méthode permet d'estimer les différents paramètres intervenants dans le modèle de Mack.

3.3 La notion de *tail factor*

Classiquement on suppose que la première ligne de notre triangle est close : il n'y a plus de sinistres ouverts

Cette hypothèse est trop forte pour les branches à déroulement long. On supposera ([18]) qu'il existe $\lambda_\infty > 1$ tel que

$$C_{i,\infty} = C_{i,n} \times \lambda_\infty.$$

Une méthode qui a souvent été utilisée a reposé sur l'idée que l'on pouvait projeter les λ_i par une **extrapolation exponentielle** (ou une extrapolation linéaire des $\log(\lambda_k - 1)$), puis on pose

$$\lambda_\infty = \prod_{k \geq n} \hat{\lambda}_k$$

Exemple 1

Sur notre triangle, cette méthode prévoit de rajouter 0.07% de charge par rapport à la prédiction faite par les méthodes classiques (en supposant la première année close).

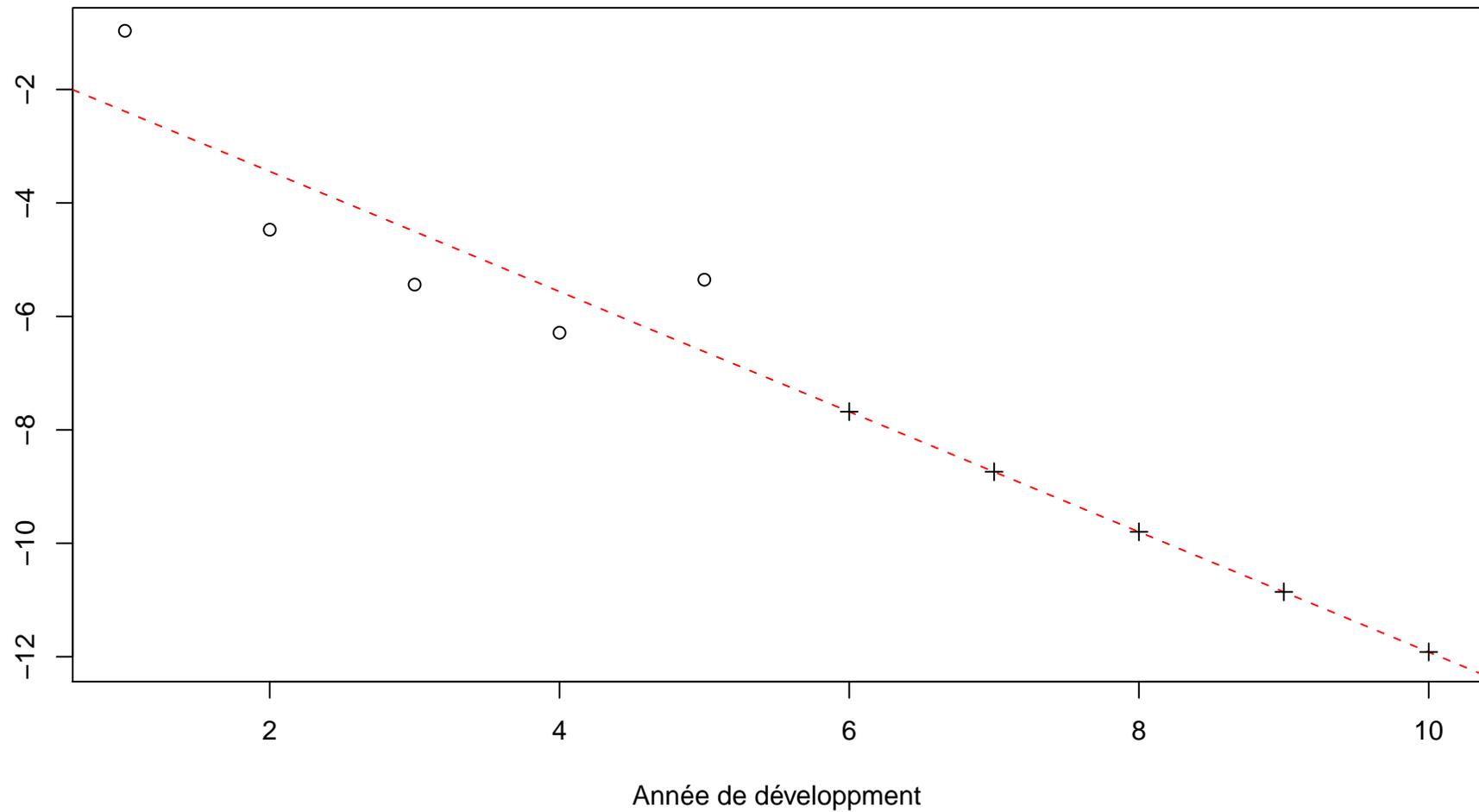


FIGURE 3 – Ajustement linéaire sur le nuage $\{(k, \log(\lambda_k - 1))\}$.

3.4 De l'incertitude sur \hat{R}_i et \hat{R}

L'incertitude est ici quantifiée à l'aide du *mean squared error*,

$$\text{mse}(\hat{R}_i) = \text{mse}(\hat{C}_{i,n} - C_{i,n-i}) = \text{mse}(\hat{C}_{i,n}) = \mathbb{E} \left(\left[\hat{C}_{i,n} - C_{i,n} \right]^2 \mid \mathcal{H}_n \right)$$

Proposition 2

Le mean squared error du montant de provision $\text{mse}(\hat{R}_i)$, pour une année de survenance i , peut être estimé par

$$\widehat{\text{mse}}(\hat{R}_i) = \hat{C}_{i,n}^2 \sum_{k=n-i}^{n-1} \frac{\hat{\sigma}_k^2}{\hat{\lambda}_k^2} \left(\frac{1}{\hat{C}_{i,k}} + \frac{1}{\sum_{j=1}^{n-k} C_{j,k}} \right).$$

Toutefois, une compagnie doit au minimum provisionner pour la branche d'activité, et par par année. Il faut ensuite calculer le mse pour

$\widehat{R} = \widehat{R}_1 + \cdots + \widehat{R}_n$. En fait, on notera que

$$\text{mse}(\widehat{R}) = \mathbb{E} \left(\left[\sum_{i=2}^n \widehat{R}_i - \sum_{i=2}^n R_i \right]^2 \middle| \mathcal{H}_n \right)$$

Proposition 3

Le mean squared error du montant de provision $\text{mse}(\widehat{R})$, pour l'ensemble des années de survenance, peut être estimé par

$$\widehat{\text{mse}}(\widehat{R}) = \sum_{i=2}^n \widehat{\text{mse}}(\widehat{R}_i) + 2 \sum_{2 \leq i < j \leq n} \widehat{C}_{i,n} \widehat{C}_{j,n} \sum_{k=n-i}^{n-1} \frac{\widehat{\sigma}_k^2 / \widehat{\lambda}_k^2}{\sum_{l=1}^{n-k} C_{l,k}}.$$

Cette vision est parfois appelée *vision à l'ultime* de l'incertitude relative au montant de provision.

Exemple 2

Sur le triangle 2 $\widehat{\text{mse}}(\widehat{R}) = 79.30$, alors que $\widehat{\text{mse}}(\widehat{R}_n) = 68.45$, $\widehat{\text{mse}}(\widehat{R}_{n-1}) = 31.3$ ou $\widehat{\text{mse}}(\widehat{R}_{n-2}) = 5.05$.

3.5 L'incertitude à un an de Merz & Wüthrich

Pour comprendre la notion d'incertitude à un an, plaçons nous un an en arrière. A la fin de l'année $n - 1$, nous disposons du triangle sans la dernière diagonale, que l'on avait alors complété par la méthode Chain Ladder (Tableau 3.5). Si l'on ne s'intéresse qu'aux années antérieures, $i = 1, \dots, n - 1$, à la fin de l'année n , nous avons obtenu un triangle avec une diagonale supplémentaire que l'on avait alors complété par la méthode Chain Ladder (Tableau 3.5).

| | 0 | 1 | 2 | 3 | 4 |
|---|------|--------|--------|--------|--------|
| 1 | 3209 | 4372 | 4411 | 4428 | 4435 |
| 2 | 3367 | 4659 | 4696 | 4720 | 4727.4 |
| 3 | 3871 | 5345 | 5398 | 5422.3 | 5430.9 |
| 4 | 4239 | 5917 | 5970.0 | 5996.9 | 6006.4 |
| 5 | 4929 | 6810.8 | 6871.9 | 6902.9 | 6939.0 |

TABLE 8 – Triangle des paiements cumulés sur les années antérieures, $\mathbf{C} = (C_{i,j})_{i+j \leq n-1, i \leq n-1}$ avec les projection future $\hat{\mathbf{C}} = (\hat{C}_{i,j})_{i+j > n-1}$.

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|------|------|--------|---------|--------|--------|
| 1 | 3209 | 4372 | 4411 | 4428 | 4435 | 4456 |
| 2 | 3367 | 4659 | 4696 | 4720 | 4730 | 4752.4 |
| 3 | 3871 | 5345 | 5398 | 5420 | 5430.1 | 5455.8 |
| 4 | 4239 | 5917 | 6020 | 6046.15 | 6057.4 | 6086.1 |
| 5 | 4929 | 6794 | 6871.7 | 6901.5 | 6914.3 | 6947.1 |

TABLE 9 – Triangle des paiements cumulés sur les années antérieures, $\mathbf{C} = (C_{i,j})_{i+j \leq n, i \leq n-1}$ avec les projection future $\hat{\mathbf{C}} = (\hat{C}_{i,j})_{i+j > n}$.

A la fin de l'année $n - 1$, le montant de provisions constitué était de 2114,61, pour ces $n - 1$ premières années. Au final, on pensait payer 27513,61 (toutes années confondues). A la fin de l'année n , la charge totale était revue à la hausse, passant à 27697,33. Cette augmentation de 183,72 est appelée *mali*. C'est l'incertitude associée à cet quantité qui est aujourd'hui demandé dans Solvabilité II.

Formellement, il convient d'introduire dans les notations la date à laquelle est faite l'estimation. Par exemple, on distinguera

$$\widehat{\lambda}_j^n = \frac{\sum_{i=1}^{n-i-1} C_{i,j+1}}{\sum_{i=1}^{n-i-1} C_{i,j}} \quad \text{et} \quad \widehat{\lambda}_j^{n+1} = \frac{\sum_{i=1}^{n-i} C_{i,j+1}}{\sum_{i=1}^{n-i} C_{i,j}}$$

La section précédente permet de montrer que

$$\mathbb{E}(\widehat{\lambda}_j^n | \mathcal{H}_n) = \lambda_j \quad \text{et} \quad \mathbb{E}(\widehat{\lambda}_j^{n+1} | \mathcal{H}_{n+1}) = \lambda_j.$$

Sauf qu'ici, on se place toujours à la date n . Il convient alors de calculer $\mathbb{E}(\widehat{\lambda}_j^{n+1} | \mathcal{H}_n)$

Notons que si l'on pose $S_j^n = C_{1,j} + C_{2,j} + \dots + C_{j,n-j}$,

$$\widehat{\lambda}_j^{n+1} = \frac{\sum_{i=1}^{n-i} C_{i,j+1}}{\sum_{i=1}^{n-i} C_{i,j}} = \frac{\sum_{i=1}^{n-i} C_{i,j+1}}{S_j^{n+1}} = \frac{\sum_{i=1}^{n-1-i} C_{i,j+1}}{S_j^{n+1}} + \frac{C_{n-j,j+1}}{S_j^{n+1}}$$

soit simplement

$$\widehat{\lambda}_j^{n+1} = \frac{S_j^n \cdot \widehat{\lambda}_j^n}{S_j^{n+1}} + \frac{C_{n-j,j+1}}{S_j^{n+1}}.$$

Lemme 1

Sous les hypothèses du modèles de Mack,

$$\mathbb{E}(\widehat{\lambda}_j^{n+1} | \mathcal{H}_n) = \frac{S_j^n}{S_j^{n+1}} \cdot \widehat{\lambda}_j^n + \lambda_j \cdot \frac{C_{n-j,n}}{S_j^{n+1}}.$$

On en déduit en particulier que

$$\mathbb{E}(\widehat{C}_{i,j}^{n+1} | \mathcal{H}_n) = C_{i,n-i} \cdot \lambda_{n-i} \cdot \prod_{k=n-i+1}^{j-1} \mathbb{E}(\widehat{\lambda}_k^{n+1} | \mathcal{H}_n).$$

En reprenant les notations de [19], on peut étudier la variation du boni/mali d'une année sur l'autre, c'est à dire du changement dans la prédiction de la charge totale, entre deux années.

Pour cela, on introduit le concept suivant

Définition 1

Le *claims development result* $CDR_i(n+1)$, pour l'année de survenance i , entre les dates n et $n+1$, aussi appelé *boni* ou *mali* (suivant qu'il est positif ou négatif), est

$$CDR_i(n+1) = \mathbb{E}(R_i^n | \mathcal{H}_n) - [Y_{i,n-i+1} + \mathbb{E}(R_i^{n+1} | \mathcal{H}_{n+1})],$$

où $Y_{i,n-i+1}$ correspond à l'incrément de paiements, $Y_{i,n-i+1} = C_{i,n-i+1} - C_{i,n-i}$.

On notera que $CDR_i(n+1)$ est une martingale \mathcal{H}_{n+1} -mesurable, et que l'on peut réécrire

$$CDR_i(n+1) = \mathbb{E}(C_{i,n} | \mathcal{H}_n) - \mathbb{E}(C_{i,n} | \mathcal{H}_{n+1}).$$

En transpirant un peu, on arrive finalement à la propriété suivante

Lemme2

Sous les hypothèses du modèle de Mack, un estimateur de $\mathbb{E}(\widehat{CDR}_i(n+1)^2 | \mathcal{H}_n)$

$$\widehat{\text{mse}}(\widehat{CDR}_i(n+1) | \mathcal{H}_n) = \widehat{C}_{i,n}^2 \left(\widehat{\Gamma}_{i,n} + \widehat{\Delta}_{i,n} \right)$$

où

$$\widehat{\Delta}_{i,n} = \frac{\widehat{\sigma}_{n-i+1}^2}{\widehat{\lambda}_{n-i+1}^2 S_{n-i+1}^{n+1}} + \sum_{j=n-i+2}^{n-1} \left(\frac{C_{n-j+1,j}}{S_j^{n+1}} \right)^2 \frac{\widehat{\sigma}_j^2}{\widehat{\lambda}_j^2 S_j^n}$$

et

$$\widehat{\Gamma}_{i,n} = \left(1 + \frac{\widehat{\sigma}_{n-i+1}^2}{\widehat{\lambda}_{n-i+1}^2 C_{i,n-i+1}} \right) \prod_{j=n-i+2}^{n-1} \left(1 + \frac{\widehat{\sigma}_j^2}{\widehat{\lambda}_j^2 [S_j^{n+1}]^2} C_{n-j+1,j} \right) - 1$$

Remarque2

On peut noter que $\widehat{C}_{i,n}^2 \widehat{\Delta}_{i,n}$ est un estimateur de

$$\mathbb{E} \left([\mathbb{E}(\widehat{CDR}_i(n+1) | \mathcal{H}_n)]^2 | \mathcal{H}_n \right).$$

En revanche $\widehat{C}_{i,n}^2 \widehat{\Gamma}_{i,n}$ est un estimateur de

$$\text{Var}(\widehat{CDR}_i(n+1) | \mathcal{H}_n).$$

[19] ont alors approché ce dernier terme terme par

$$\widehat{\Gamma}_{i,n} \approx \frac{\widehat{\sigma}_{n-i+1}^2}{\widehat{\lambda}_{n-i+1}^2 C_{i,n-i+1}} + \sum_{j=n-i+2}^{n-1} \left(\frac{C_{n-j+1,j}}{S_j^{n+1}} \right)^2 \frac{\widehat{\sigma}_j^2}{\widehat{\lambda}_j^2 C_{n-j+1,j}}$$

en faisant tout simplement un développement de la forme $\prod(1 + u_i) \approx 1 + \sum u_i$, mais qui n'est valide que si u_i est petit, soit ici

$$\frac{\widehat{\sigma}_j^2}{\widehat{\lambda}_j^2} \ll C_{n-j+1,j}.$$

Pour l'erreur d'estimation, on parle également d'*erreur de prédiction rétrospective*, il convient de calculer, puis d'estimer

$$\mathbb{E} \left([CDR_i(n+1) - \widehat{CDR}_i(n+1)]^2 | \mathcal{H}_n \right),$$

et pas seulement

$$\mathbb{E}(\widehat{CDR}_i(n+1)^2 | \mathcal{H}_n)$$

comme nous venons de le faire. Mais comme nous le rappellerons par la suite, la réglementation ne tient pas compte de ce terme, car il est illégal de supposer $\widehat{CDR}_i(n+1) \neq 0$.

Toutefois, pour ce terme, il est possible de montrer la propriété suivante

Lemme3

Sous les hypothèses du modèle de Mack, un estimateur de

$\mathbb{E}([\widehat{CDR}_i(n+1) - CDR_i(n+1)]^2 | \mathcal{H}_n)$ est

$$\widehat{\text{mse}}(CDR_i(n+1) | \mathcal{H}_n) = \widehat{C}_{i,n}^2 \left(\widehat{\Phi}_{i,n} + \widehat{\Delta}_{i,n} \right)$$

où

$$\widehat{\Phi}_{i,n} = \left(1 + \frac{\widehat{\sigma}_{n-i+1}^2}{\widehat{\lambda}_{n-i+1}^2 C_{i,n-i+1}} \right) \left(\prod_{j=n-i+2}^{n-1} \left(1 + \frac{\widehat{\sigma}_j^2}{\widehat{\lambda}_j^2 [S_j^{n+1}]^2} C_{n-j+1,j} \right) - 1 \right).$$

[19] ont alors approché ce dernier terme terme par

$$\widehat{\Phi}_{i,n} \approx \sum_{j=n-i+2}^{n-1} \frac{[\widehat{\sigma}_j^n]^2}{[\widehat{\lambda}_j^n]^2 C_{n-j+1,j}} \left(\frac{C_{n-j+1,j}}{S_j^{n+1}} \right)^2,$$

en faisant là encore un développement de la forme $\prod(1 + u_i) \approx 1 + \sum u_i$, mais qui n'est valide que si u_i est petit, soit ici

$$\frac{\widehat{\sigma}_j^2}{\widehat{\lambda}_j^2} \ll C_{n-j+1,j}.$$

En considérant ces approximations, et en supposant qu'il est légalement impossible de prédire autre chose que $\widehat{CDR}_i(n+1) = 0$, on peut finalement réécrire

Proposition 4

Sous les hypothèses du modèle de Mack,

$$\widehat{\text{mse}}_n(\widehat{CDR}_i(n+1)) \approx [\widehat{C}_{i,n}^n]^2 \left[\frac{[\widehat{\sigma}_{n-i+1}^n]^2}{[\widehat{\lambda}_{n-i+1}^n]^2} \left(\frac{1}{\widehat{C}_{i,n-i+1}} + \frac{1}{\widehat{S}_{n-i+1}^n} \right) + \sum_{j=n-i+2}^{n-1} \frac{[\widehat{\sigma}_j^n]^2}{[\widehat{\lambda}_j^n]^2} \left(\frac{1}{\widehat{S}_j^n} \left(\frac{\widehat{C}_{n-j+1,j}}{\widehat{S}_j^{n+1}} \right)^2 \right) \right].$$

Pour rappel, la formule de Mack proposait

$$\widehat{\text{mse}}_n(\widehat{R}_i) = [\widehat{C}_{i,n}^n]^2 \left[\frac{[\widehat{\sigma}_{n-i+1}^n]^2}{[\widehat{\lambda}_{n-i+1}^n]^2} \left(\frac{1}{\widehat{C}_{i,n-i+1}} + \frac{1}{\widehat{S}_{n-i+1}^n} \right) + \sum_{j=n-i+2}^{n-1} \frac{[\widehat{\sigma}_j^n]^2}{[\widehat{\lambda}_j^n]^2} \left(\frac{1}{\widehat{C}_{i,j}} + \frac{1}{\widehat{S}_j^n} \right) \right].$$

Autrement dit, dans le mse du CDR, seulement le première terme de l'erreur de modèle de la formule de Mack est considérée, et pour l'erreur d'estimation, on ne considère que la première diagonale $i + j = n + 1$ (les termes suivants étant écrasés par le facteur $\widehat{C}_{n-j+1,j} / \widehat{S}_j^{n+1}$).

Enfin, si l'on regarde finalement ce qui se passe toutes années de survenance confondues, on a une formule qui peut encore se rapprocher de celle proposée par Mack, à savoir

$$\widehat{\text{mse}}_n(\text{CDR}(n+1)) \approx \sum_{i=1}^n \widehat{\text{mse}}_n(\text{CDR}_i(n+1)) + 2 \sum_{i < l} \widehat{C}_{i,n}^n \widehat{C}_{l,n}^n \left(\frac{[\widehat{\sigma}_{n-i}^n]^2 / [\widehat{\lambda}_{n-i}^n]^2}{\sum_{k=0}^{i-1} C_{k,n-i}} + \sum_{j=n-i+1}^{n-1} \frac{C_{n-j,j}}{\sum_{k=0}^{n-j} C_{k,j}} \frac{[\widehat{\sigma}_j^n]^2 / [\widehat{\lambda}_j^n]^2}{\sum_{k=0}^{n-j-1} C_{k,j}} \right).$$

Cette approximation n'est toutefois valide que si $C_{n-j+1,j} \leq S_j^{n+1}$.

Exemple3

Sur le triangle 2 $\widehat{\text{mse}}_n(\text{CDR}(n+1)) = 72.57$, alors que $\widehat{\text{mse}}_n(\text{CDR}_n(n+1)) = 60.83$, $\widehat{\text{mse}}_n(\text{CDR}_{n-1}(n+1)) = 30.92$ ou encore $\widehat{\text{mse}}_n(\text{CDR}_{n-2}(n+1)) = 4.48$. La formule approchée donne des résultats semblables.

4 Régression Poissonnienne et approches économétriques

Dans cette section, nous nous éloignerons des modèles récurrents inspirés de la méthode Chain Ladder, et nous reviendrons sur des classes de modèles très utilisés dans les années 70, appelés *modèles à facteurs*, remis au goût du jour en proposant une lecture économétrique de ces modèles, permettant ainsi d'obtenir des intervalles de confiance des différentes grandeurs.

4.1 Les modèles à facteurs, un introduction historique

Avant de présenter l'utilisation des modèles de régression, on peut commencer par évoquer des modèles plus anciens. Par exemple [26] supposait que

$$Y_{i,j} = r_j \cdot \mu_{i+j}, \text{ pour tout } i, j$$

i.e. le paiement effectué l'année $i + j$ pour les sinistres survenus l'année i est fonction d'un effet colonne (de cadence de paiement) et un effet diagonal, que Taylor interprète comme un facteur d'inflation. Ce modèle peut se réécrire, **dès lors qu'il n'y a pas d'incrément positif**,

$$\log Y_{i,j} = \alpha_i + \gamma_{i+j}$$

qui prend alors une forme linéaire.

Comme nous l'avons noté un modèle de la forme

$$Y_{i,j} = a_i \times b_j$$

pourra se rapprocher du modèle Chain Ladder.

4.2 Les modèles Gaussiens de de Vylder et de Christophides

Considèrerons un modèle

$$Y_{i,j} \sim \mathcal{N}(\alpha_i \beta_j, \sigma^2), \forall i, j$$

On peut estimer les coefficients par moindres carrés,

$$(\hat{\alpha}, \hat{\beta}) = \operatorname{argmin} \left\{ \sum_{i,j} [Y_{i,j} - \alpha_i \beta_j]^2 \right\}.$$

Les équations normales s'écrivent ici $\hat{\alpha}_i = \frac{\sum_j Y_{i,j} \hat{\beta}_j}{\sum_j \hat{\beta}_j^2}$ et $\hat{\beta}_j = \frac{\sum_i Y_{i,j} \hat{\alpha}_i}{\sum_i \hat{\alpha}_i^2}$, ce qui ne résoud pas explicitement. Pour le résoudre, [4] a suggéré de le réécrire comme un modèle log-linéaire, i.e.

$$\log Y_{i,j} \sim \mathcal{N}(a_i + b_j, \sigma^2), \forall i, j$$

4.3 La régression poissonnienne de Hachemeister & Stanard

[10], [13] et enfin [16] ont montré que dans une *régression log-Poisson* sur les incréments, la somme des prédictions des paiements à venir correspond à l'estimateur Chain Ladder.

On suppose ici que $\mathbb{E}(Y_{i,j}) = \mu_{i,j} = \exp[r_i + c_j] = a_i \cdot b_j$.

Il y a ici un $2n$ paramètres à estimer, $\mathbf{a} = (a_0, \dots, a_n)$ et $\mathbf{b} = (b_0, \dots, b_n)$, avec une contrainte de la forme $b_0 + \dots + b_n = 1$ (car il nous reste un degré de liberté). C

une fois estimés ces paramètres, il est possible de *prédire* la partie inférieure du triangle très simplement, i.e.

$$\hat{Y}_{i,j} = \hat{\mu}_{i,j} = \exp[\hat{r}_i + \hat{c}_j] = \hat{a}_i \cdot \hat{b}_j.$$

Exemple4

Sur le triangle 2, on obtient la sortie de régression suivante, en régressant sur un facteur

ligne, et un facteur colonne.

Call:

```
glm(formula = Y ~ lig + col, family = poisson("log"), data = base)
```

Coefficients:

| | Estimate | Std. Error | z value | Pr(> z) | |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 8.05697 | 0.01551 | 519.426 | < 2e-16 | *** |
| lig2 | 0.06440 | 0.02090 | 3.081 | 0.00206 | ** |
| lig3 | 0.20242 | 0.02025 | 9.995 | < 2e-16 | *** |
| lig4 | 0.31175 | 0.01980 | 15.744 | < 2e-16 | *** |
| lig5 | 0.44407 | 0.01933 | 22.971 | < 2e-16 | *** |
| lig6 | 0.50271 | 0.02079 | 24.179 | < 2e-16 | *** |
| col2 | -0.96513 | 0.01359 | -70.994 | < 2e-16 | *** |
| col3 | -4.14853 | 0.06613 | -62.729 | < 2e-16 | *** |
| col4 | -5.10499 | 0.12632 | -40.413 | < 2e-16 | *** |
| col5 | -5.94962 | 0.24279 | -24.505 | < 2e-16 | *** |
| col6 | -5.01244 | 0.21877 | -22.912 | < 2e-16 | *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 46695.269 on 20 degrees of freedom
Residual deviance: 30.214 on 10 degrees of freedom
(15 observations deleted due to missingness)
AIC: 209.52

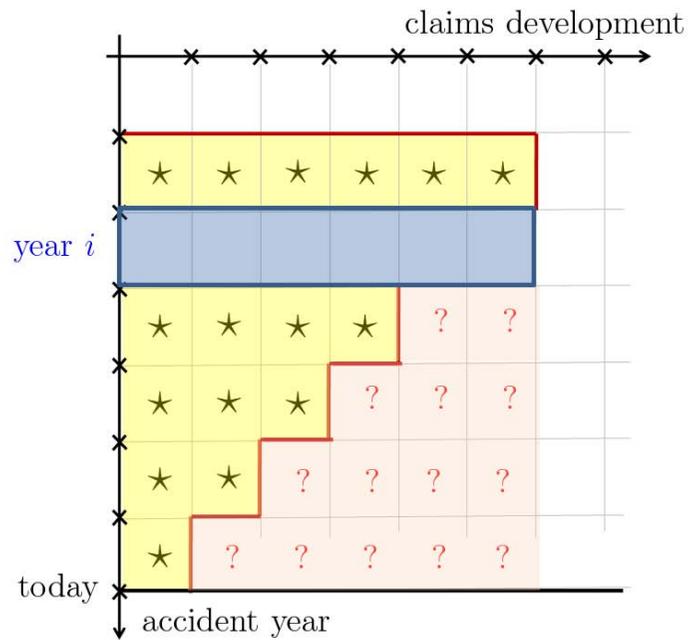
Number of Fisher Scoring iterations: 4

Les prédictions $\hat{Y}_{i,j}$ sont indiquées dans le tableau 4.3.

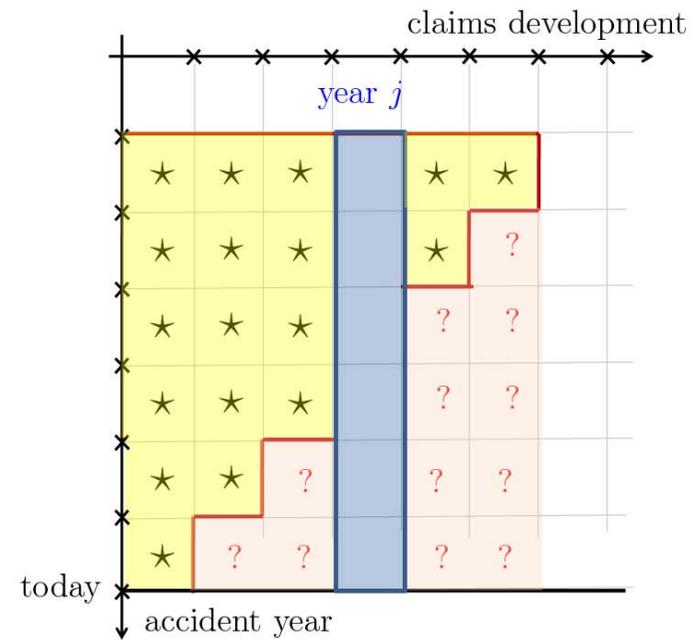
| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|------|--------|--------|---------|--------|--------|
| 1 | 3209 | 4372 | 4411 | 4428 | 4435 | 4456 |
| 2 | 3367 | 4659 | 4696 | 4720 | 4730 | 4752.4 |
| 3 | 3871 | 5345 | 5398 | 5420 | 5430.1 | 5455.8 |
| 4 | 4239 | 5917 | 6020 | 6046.15 | 6057.4 | 6086.1 |
| 5 | 4929 | 6794 | 6871.7 | 6901.5 | 6914.3 | 6947.1 |
| 6 | 5217 | 7204.3 | 7286.7 | 7318.3 | 7331.9 | 7366.7 |

TABLE 10 – Triangle des prédictions d’incrément de paiements, $\hat{\mathbf{Y}} = (\hat{Y}_{i,j})_{0 \leq i,j \leq n}$ par une régression log-Poisson.

the occurrence factor α_i

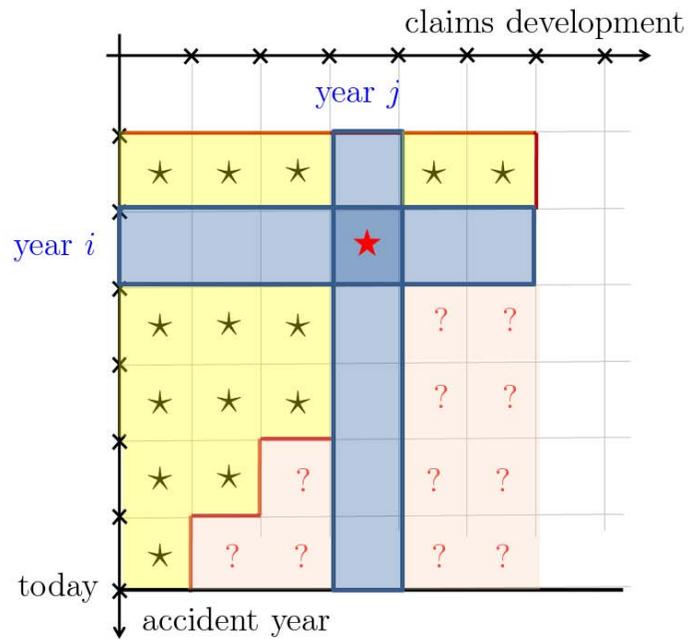


the development factor β_j



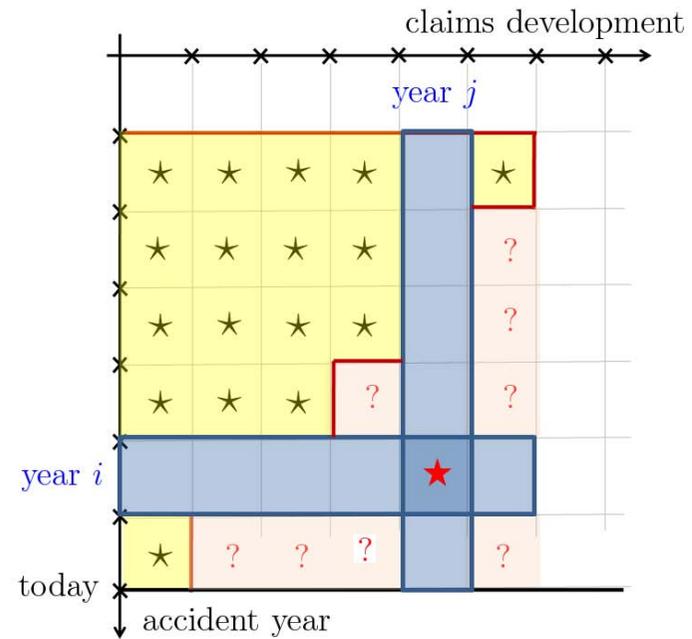
$$\widehat{Y}_{i,j} = \exp[\widehat{\alpha}_i + \widehat{\beta}_j]$$

on past observations



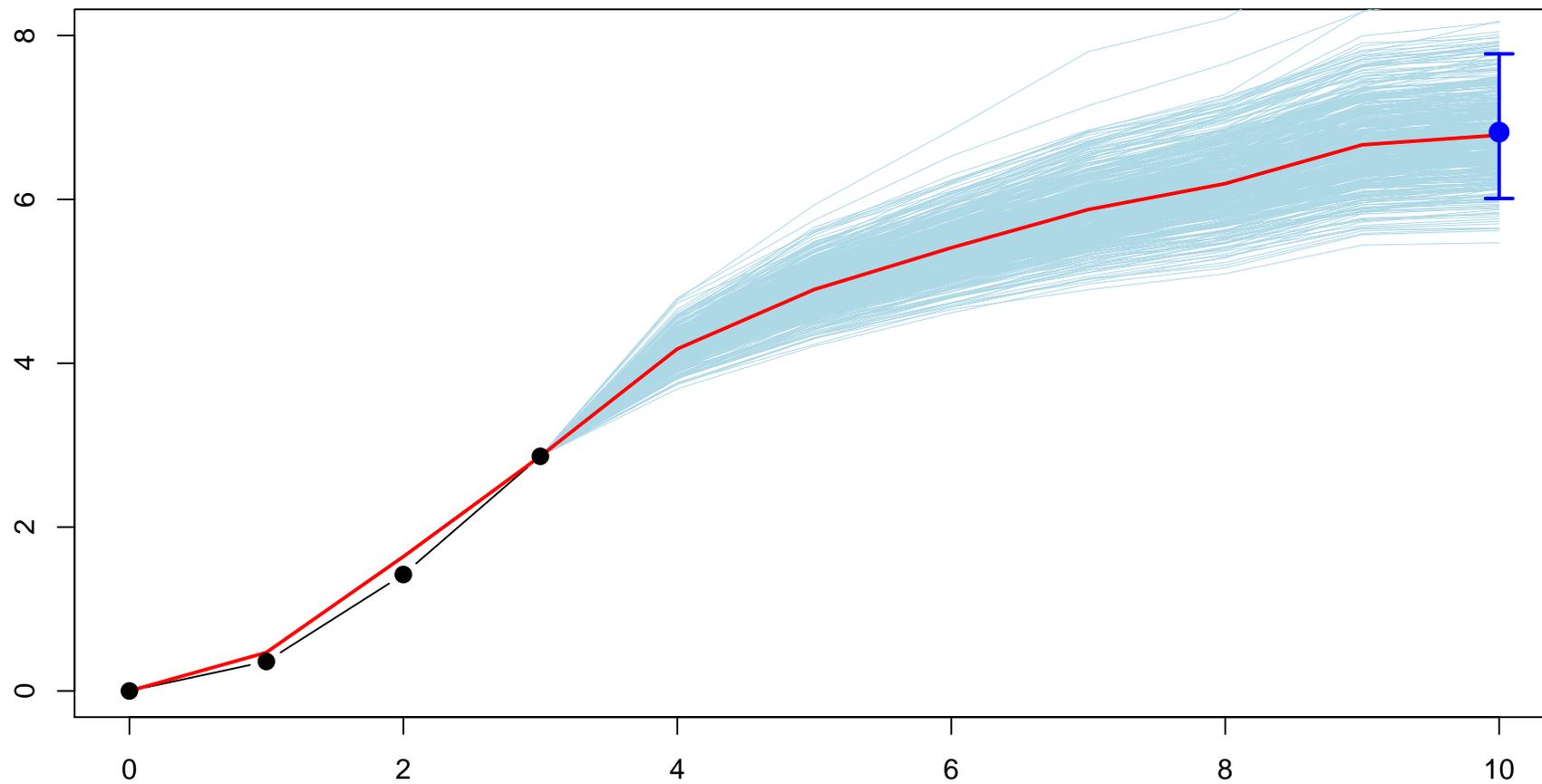
$$\widehat{Y}_{i,j} = \exp[\widehat{\alpha}_i + \widehat{\beta}_j]$$

on the future



4.4 Incertitude dans un modèle de régression

Nous avons noté auparavant qu'obtenir une estimation du montant de sinistres restant à payer ne suffisait pas, et qu'il fallait avoir un intervalle de confiance, ou - au moins - une mesure de la dispersion du vrai montant autour de cette valeur prédite.



4.4.1 Les formules économétriques fermées

Les modèles de régressions pourraient paraître très intéressants car il existe des formules fermés pour toutes sortes de prédictions. Par exemple, dans une **régression GLM avec un lien logarithmique**, rappelons que

$$\mathbb{E}(Y_{i,j} | \mathcal{H}_n) = \mu_{i,j} = \exp[\eta_{i,j}]$$

ou encore

$$\widehat{Y}_{i,j} = \widehat{\mu}_{i,j} = \exp[\widehat{\eta}_{i,j}].$$

La **delta method** nous permet d'écrire que

$$\text{Var}(\widehat{Y}_{i,j}) \approx \left| \frac{\partial \mu_{i,j}}{\partial \eta_{i,j}} \right|^2 \cdot \text{Var}(\widehat{\eta}_{i,j}),$$

ce qui se simplifie dans le cas où le lien est logarithmique, i.e.

$$\frac{\partial \mu_{i,j}}{\partial \eta_{i,j}} = \mu_{i,j}$$

Aussi, pour une loi de Poisson surdispersée (comme dans [20]),

$$\mathbb{E} \left([Y_{i,j} - \widehat{Y}_{i,j}]^2 \right) \approx \widehat{\phi} \cdot \widehat{\mu}_{i,j} + \widehat{\mu}_{i,j}^2 \cdot \widehat{\text{Var}}(\widehat{\eta}_{i,j})$$

pour la partie inférieure du triangle. Et en notant que

$$\text{Cov}(\widehat{Y}_{i,j}, \widehat{Y}_{k,l}) \approx \widehat{\mu}_{i,j} \cdot \widehat{\mu}_{k,l} \cdot \widehat{\text{Cov}}(\widehat{\eta}_{i,j}, \widehat{\eta}_{k,l}).$$

on peut calculer

$$\mathbb{E} \left([R - \widehat{R}]^2 \right) \approx \left(\sum_{i+j>n} \widehat{\phi} \cdot \widehat{\mu}_{i,j} \right) + \widehat{\mu}' \cdot \widehat{\text{Var}}(\widehat{\eta}) \cdot \widehat{\mu}$$

Remarque3

Cette formule est malheureusement asymptotique, ce qui est rarement le cas en provisionnement où l'on dispose de très peu de données.

Exemple5

Sur notre triangle, on obtient un mean squared error de l'ordre de 131.77.

4.4.2 Les méthodes de simulations

Les méthodes de simulation sont une bonne alternative si on dispose de trop peu de données pour invoquer des théorèmes asymptotiques. Rappelons, comme le notait [17] qu'il existe 2 sources d'incertitude,

- l'erreur de modèle (on parle de *process error*)
- l'erreur d'estimation (on parle de *variance error*)

Il sera alors nécessaire d'utiliser deux algorithmes pour quantifier ces deux erreurs.

Afin de quantifier l'erreur d'estimation, il est naturel de simuler des faux triangles (supérieurs), puis de regarder la distribution des estimateurs de montant de provisions obtenus pour chaque triangles. A l'étape b , on génère un pseudo triangle à l'aide des résidus de Pearson. Rappelons que pour une régression de Poisson,

$$\hat{\varepsilon}_{i,j} = \frac{Y_{i,j} - \hat{\mu}_{i,j}}{\sqrt{\hat{\mu}_{i,j}}}.$$

Les erreurs de Pearson obtenues peuvent être visualisées dans le Tableau 4.4.2

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|--------|--------|--------|--------|--------|-------|
| 1 | 0.948 | -1.128 | -1.533 | -0.489 | -0.427 | 0.000 |
| 2 | 0.024 | 0.277 | -2.213 | 0.792 | 0.414 | |
| 3 | 0.116 | 0.056 | -1.024 | -0.297 | | |
| 4 | -1.082 | 0.891 | 4.237 | | | |
| 5 | 0.130 | -0.211 | | | | |
| 6 | 0.000 | | | | | |

TABLE 11 – Le triangle des résidus de Pearson, $\hat{\varepsilon}$ où $\hat{\varepsilon}_{i,j} = \hat{\mu}_{i,j}^{-1/2} \cdot [Y_{i,j} - \hat{\mu}_{i,j}]$.

On considère les erreurs de Pearson de manière à *normaliser* les résidus (que l'on pourra alors supposer i.i.d).

Remarque4

Les résidus de Pearson ne sont généralement pas Gaussien.

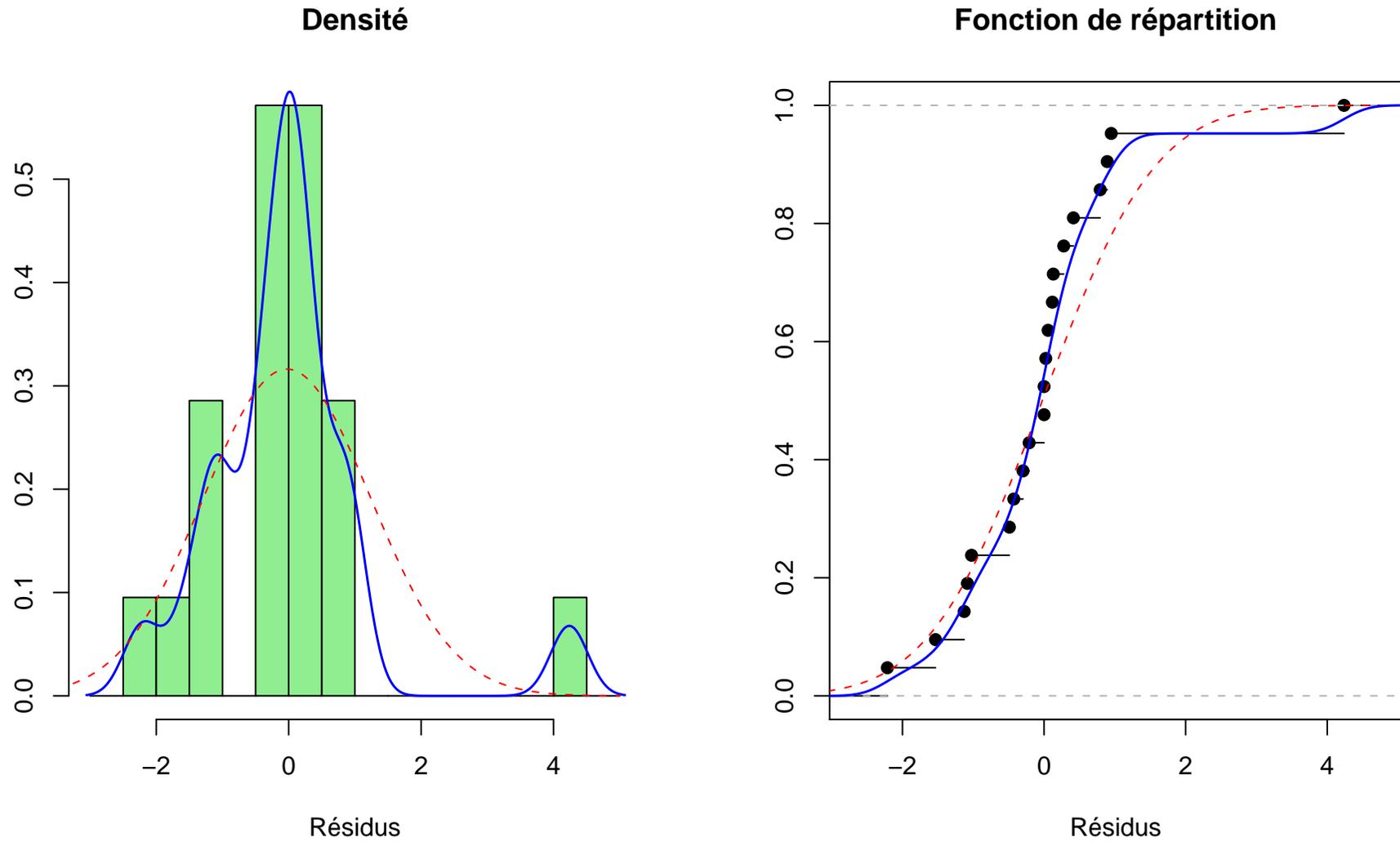


FIGURE 4 – Distribution empirique des résidus de Pearson.

En simulant des erreurs (qui sont supposées indépendantes et identiquement distribuée), $\tilde{\varepsilon}^b = (\tilde{\varepsilon}_{i,j}^b)$, on pose alors

$$Y_{i,j}^b = \hat{\mu}_{i,j} + \sqrt{\hat{\mu}_{i,j}} \cdot \tilde{\varepsilon}_{i,j}^b.$$

Pour générer des erreurs, la méthode la plus usuelle est d'utiliser une simulation nonparamétrique, c'est à dire que l'on va **bootstrapper les résidus** parmi les pseudorésidus obtenus $\hat{\varepsilon}$. Sinon il est aussi possible d'utiliser un **modèle paramétrique** (par exemple supposer une loi normale, même si rien - théoriquement - ne justifie cette méthode).

Le Tableau 4.4.2 présente un triangle simulé $\mathbf{Y}^b = (Y_{i,j}^b)$.

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|--------|--------|-------|-------|------|------|
| 1 | 3155.7 | 1216.4 | 42.2 | 18.2 | 9.0 | 22.9 |
| 2 | 3381.7 | 1245.4 | 84.0 | 18.2 | 11.1 | |
| 3 | 3726.1 | 1432.5 | 61.44 | 23.44 | | |
| 4 | 4337.3 | 1642.8 | 74.6 | | | |
| 5 | 4929.0 | 1879.8 | | | | |
| 6 | 5186.1 | | | | | |

TABLE 12 – Triangle de paiements bootstrapé, $\mathbf{Y}^b = (Y_{i,j}^b)$.

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|--------|--------|--------|--------|--------|-------|
| 1 | 0.00 | 0.414 | -1.082 | -0.211 | 0.0277 | 0.414 |
| 2 | 0.277 | -1.024 | 4.237 | -0.489 | 0.792 | |
| 3 | -2.213 | -1.024 | 0.056 | 0.000 | | |
| 4 | 0.414 | 0.024 | 0.792 | | | |
| 5 | 0.130 | 0.130 | | | | |
| 6 | -0.427 | | | | | |

TABLE 13 – Triangle de résidus simulé par bootstrap, $\varepsilon^b = (\varepsilon_{i,j}^b)$.

Une fois simulé un pseudo triangle d'incrément de paiements, on prédit un montant de provision \hat{R}^b (par exemple via une méthode Chain Ladder, ou en refaisant une modélisation log-Poisson si les incréments de paiements sont tous positifs). La variance des \hat{R}^b correspond à l'erreur d'estimation.

Le Tableau 14 présente la modélisation du triangle simulé 4.4.2 par un modèle log-Poisson $\hat{\mathbf{Y}}^b = (\hat{Y}_{i,j}^b)$. La somme des termes dans la partie inférieure donne une valeur possible pour l'estimation du montant de provision $\hat{R}^b = 2448.175$.

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|--------|--------|------|------|------|------|
| 1 | 3157.5 | 1199.1 | 56.7 | 18.4 | 9.7 | 22.8 |
| 2 | 3369.9 | 1279.8 | 60.5 | 19.6 | 10.3 | 24.4 |
| 3 | 3735.8 | 1418.7 | 67.1 | 21.7 | 11.5 | 27.0 |
| 4 | 4331.7 | 1645.1 | 77.8 | 25.2 | 13.3 | 31.4 |
| 5 | 4934.6 | 1874.0 | 88.6 | 28.7 | 15.2 | 35.7 |
| 6 | 5186.1 | 1969.5 | 93.1 | 30.2 | 16.0 | 37.6 |

TABLE 14 – Triangle des prédictions $\hat{\mathbf{Y}}^b = (\hat{Y}_{i,j}^b)$ obtenues sur le triangle simulé, $\mathbf{Y}^b = (Y_{i,j}^b)$.

Afin de prendre en compte l'erreur de modèle, plusieurs méthodes peuvent être utilisées. La première, et la plus simple, consiste à noter qu'à partir du pseudo triangle $Y_{i,j}^b$, peut obtenir des prédictions pour la partie inférieure, $\hat{Y}_{i,j}^b$. Compte tenu du modèle Poissonien, on peut alors simuler une trajectoire possible d'incrément de paiements en simulant les $Y_{i,j}^b$ à l'aide de loi de Poisson de paramètre $\hat{Y}_{i,j}^b$.

Le Tableau 4.4.2 une simulation de paiements futurs à partir du triangle simulé 4.4.2 $\mathbf{Y}^b \sim \mathcal{P}(\hat{\mathbf{Y}}^b)$. La somme des termes dans la partie inférieure dans une valeur possible pour les paiements restant à faire R^b .

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|------|-----|----|----|----|
| 1 | | | | | | |
| 2 | | | | | | 29 |
| 3 | | | | | 10 | 24 |
| 4 | | | | 27 | 11 | 37 |
| 5 | | | 94 | 34 | 16 | 40 |
| 6 | | 1956 | 102 | 26 | 15 | 36 |

TABLE 15 – Triangle de paiements simulés, $\mathbf{Y}^b \sim \mathcal{P}(\hat{\mathbf{Y}}^b)$.

La seconde méthode est d'utiliser une relecture du modèle de [17], proposée par [8]. A partir du pseudo triangle, on va utiliser les facteurs de développement $\hat{\lambda}_j$ et les variances associés $\hat{\sigma}_j^2$ obtenus sur le triangle initial. On prolonge alors le

triangle dans la partie inférence via le modèle dynamique

$$\widehat{C}_{i,j+1}^b | (\widehat{C}_{i,j}^b, \dots, \widehat{C}_{i,0}^b) \sim \mathcal{N}(\widehat{\lambda}_j \widehat{C}_{i,j}^b, \widehat{\sigma}_j^2 \widehat{C}_{i,j}^b).$$

Exemple6

Sur le triangle 2 la variance empirique de R nous permet d'avoir un bon estimateur de $\text{mse}(\widehat{R})$. Ici, à l'aide de 100 000 simulations, on obtient la distribution pour R donnée sur la Figure 5. Sur cet exemple, on obtient un écart-type (empirique) pour les simulations de R^b de l'ordre de 84.12 (ce qui se rapproche davantage des 79.3 de la méthode de Mack, que des 131.7 obtenus par développements asymptotiques).

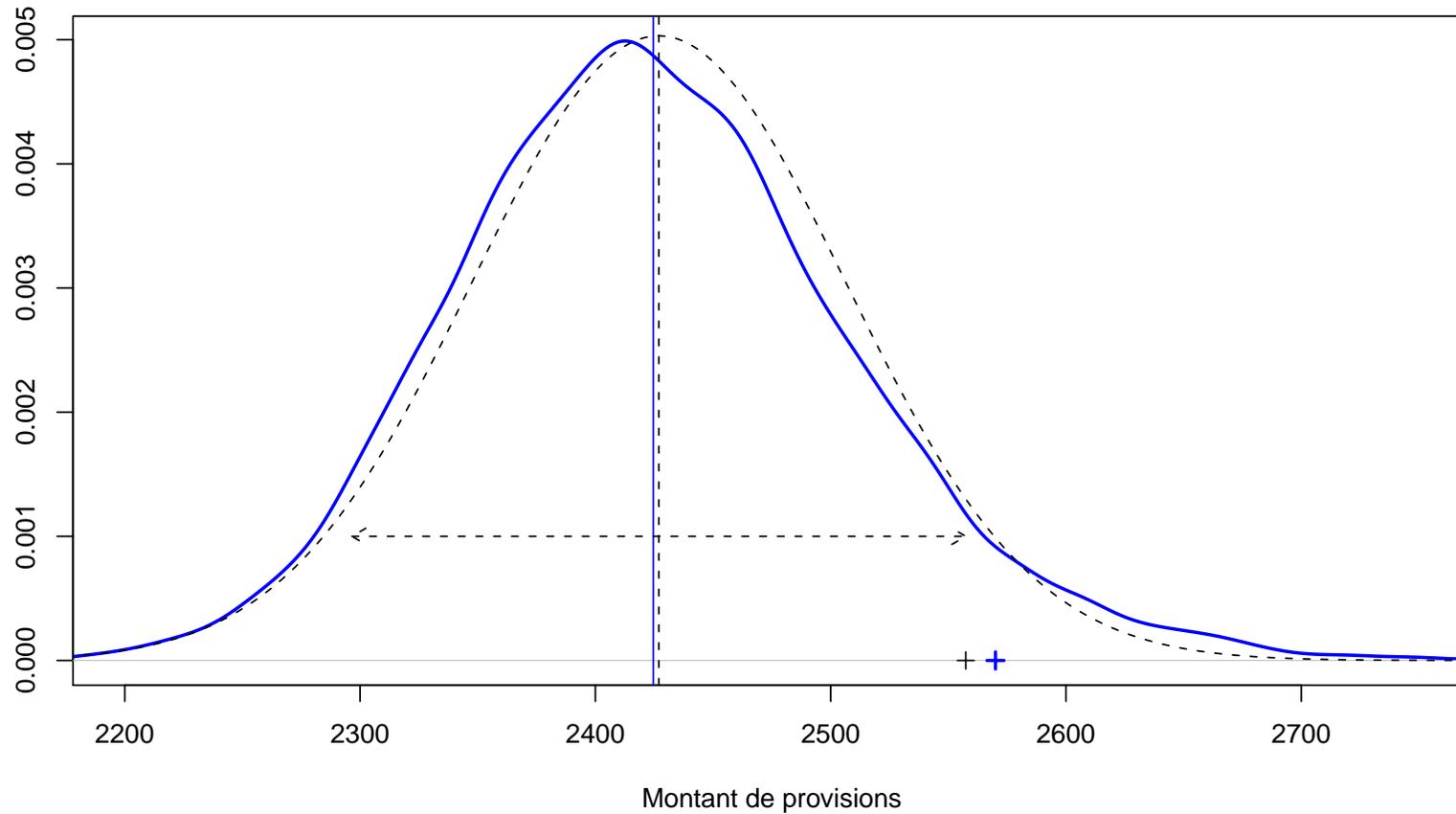


FIGURE 5 – Distribution de R , par [simulation](#) et par la méthode de Mack (sous hypothèse de normalité de R)

4.5 Quel modèle de régression ?

Nous avons justifié l'utilisation de la loi de Poisson, car le montant de provisions prédit coïncidait avec l'estimateur Chain Ladder. Mais rien ne permet de justifier statistiquement l'adéquation de la loi de Poisson a nos données.

4.5.1 Prise en compte d'une éventuelle surdispersion

En fait, la régression *quasi* Poisson peut éventuellement être plus adaptée (comme cela est mentionné par exemple dans [24]).

Exemple7

Sur le triangle 2 une modélisation par une loi quasiPoisson donne la sortie suivante

Call:

```
glm(formula = Y ~ lig + col, family = quasipoisson("log"), data = base)
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
```

| | | | | | |
|-------------|----------|---------|---------|----------|-----|
| (Intercept) | 8.05697 | 0.02769 | 290.995 | < 2e-16 | *** |
| lig2 | 0.06440 | 0.03731 | 1.726 | 0.115054 | |
| lig3 | 0.20242 | 0.03615 | 5.599 | 0.000228 | *** |
| lig4 | 0.31175 | 0.03535 | 8.820 | 4.96e-06 | *** |
| lig5 | 0.44407 | 0.03451 | 12.869 | 1.51e-07 | *** |
| lig6 | 0.50271 | 0.03711 | 13.546 | 9.28e-08 | *** |
| col2 | -0.96513 | 0.02427 | -39.772 | 2.41e-12 | *** |
| col3 | -4.14853 | 0.11805 | -35.142 | 8.26e-12 | *** |
| col4 | -5.10499 | 0.22548 | -22.641 | 6.36e-10 | *** |
| col5 | -5.94962 | 0.43338 | -13.728 | 8.17e-08 | *** |
| col6 | -5.01244 | 0.39050 | -12.836 | 1.55e-07 | *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 3.18623)

Null deviance: 46695.269 on 20 degrees of freedom

Residual deviance: 30.214 on 10 degrees of freedom
(15 observations deleted due to missingness)

AIC: NA

Number of Fisher Scoring iterations: 4

Le paramètre de surdispersion $\hat{\phi}$ vaut ici 3.18 (qui est significativement supérieur à 1).

L'estimation (mais pas la significativité) donne les mêmes résultats que la régression de Poisson, toutefois, il faut alors pour *simuler* une loi quasi Poisson.

La simulation de cette *quasi* loi, se fait généralement suivant une des deux méthodes suivantes. La première idée est d'utiliser une approximation par une loi binomiale négative $BN(r, p)$. Et l'autre est d'utiliser une approximation par une loi Gamma (dont on pourra prendre la partie entière).

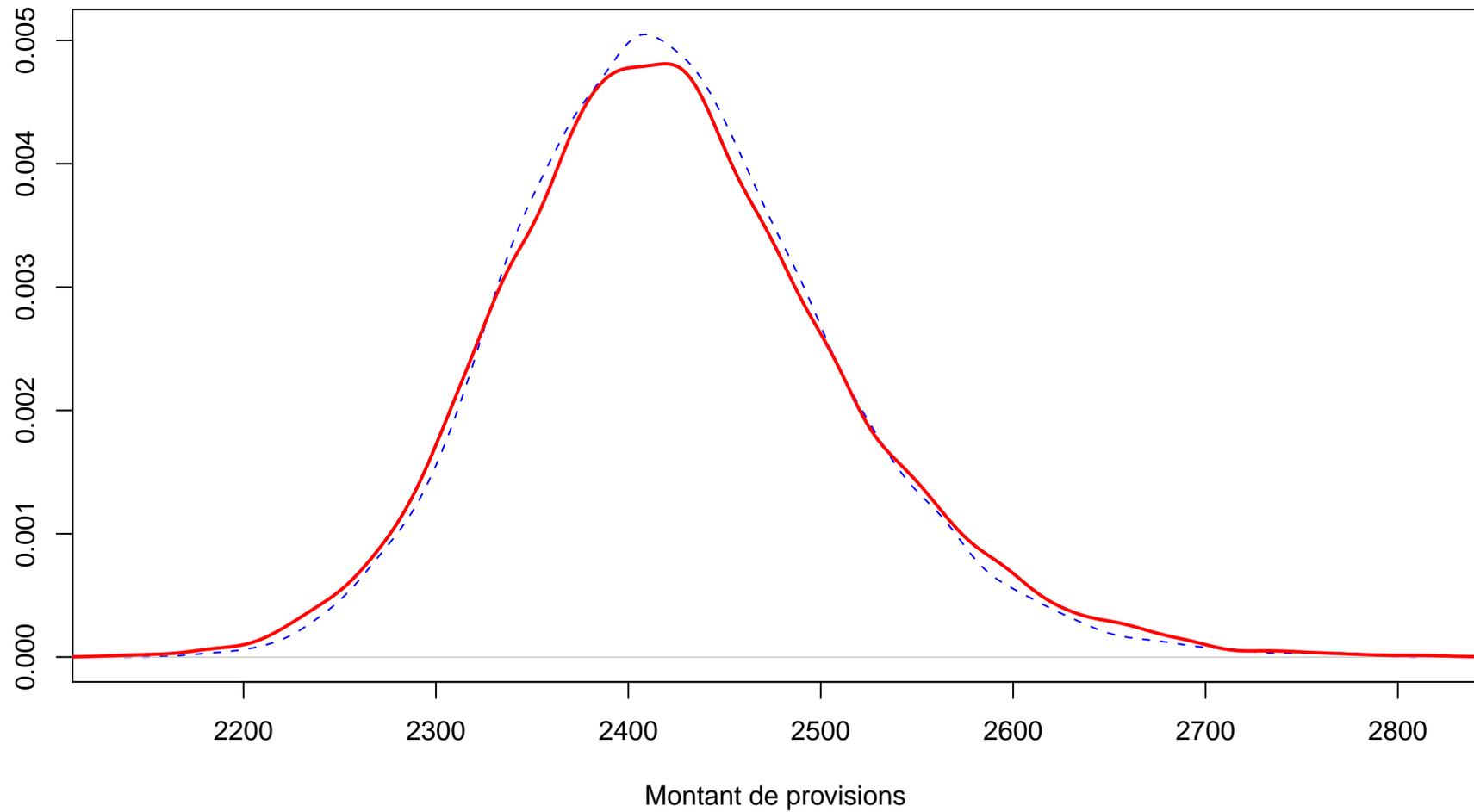


FIGURE 6 – Distribution de R , avec simulations *quasi Poisson*.

4.5.2 Entre un modèle Poisson et un modèle Gamma

Si les modèles GLM sont généralement présentés comme des modèles où une hypothèse de loi est faite (la loi de la variable dépendante Y devant appartenir à la famille exponentielle), rappelons les modèles GLM peuvent être caractérisés par une modélisation des deux premiers moments,

$$\mathbb{E}(Y|\mathbf{X}) = h(\mathbf{X}\boldsymbol{\beta}) \text{ et } \text{Var}(Y|\mathbf{X}) = V(\mathbb{E}(Y|\mathbf{X})) = g((\mathbf{X}\boldsymbol{\beta})),$$

où

– la **fonction lien**, h , qui lie la prédiction aux facteurs, ici

$$\hat{Y}_{i,j} = \mathbb{E}(Y_{i,j}|\mathcal{H}_n) = \exp[\hat{\gamma} + \hat{\alpha}_i + \hat{\beta}_j],$$

– la **loi** ou la **fonction variance**, V , qui donne la forme de la dispersion, ici

$$\text{Var}(Y_{i,j}|\mathcal{H}_n) = \phi \cdot \mathbb{E}(Y_{i,j}|\mathcal{H}_n).$$

L'unique motivation du modèle précédent (loi de Poisson, i.e. fonction variance identité, et lien logarithmique) est qu'il permet d'obtenir exactement le même montant que la méthode Chain Ladder. Mais aucun critère statistique n'a été évoqué, pour l'instant, afin de légitimer ce modèle.

Les modèles **Tweedie** sont une famille de *sur*-modèle, incluant le modèle Poissonien. On suppose que

- la **fonction lien**, est une fonction puissance, ou plutôt une transformée de Box-Cox, $\hat{Y}_{i,j} = g_{\lambda}^{-1}[\hat{\gamma} + \hat{\alpha}_i + \hat{\beta}_j]$ où $g_{\lambda}(x) = \lambda^{-1}[x^{\lambda} - 1]$ si $\lambda > 0$, avec le cas limite $g_0(x) = \log(x)$.
- la **fonction variance**, qui donne la forme de l'intervalle de confiance, ici $\text{Var}(Y_{i,j}|\mathcal{H}_n) = \phi \cdot \mathbb{E}(Y_{i,j}|\mathcal{H}_n)^{\kappa}$

où les paramètres λ et κ sont inconnus.

Remarque5

Formellement, dans les modèles GLM, on suppose que la loi de $Y_{i,j}$ appartient à la famille exponentielle, i.e. de densité

$$f(y; \theta_{i,j}, \phi) = \exp \left(\frac{y\theta_{i,j} - b(\theta_{i,j})}{\phi} + c(y, \phi) \right).$$

La fonction variance est ici $V(\mu) = b''([b']^{-1})(\mu)$. Pour reprendre les notations de [12],

on se limite ici aux formes $V(\mu) = \mu^\kappa$. Cela signifie que

$$b(\theta) = \frac{1}{2 - \kappa} [(1 - \kappa)\theta]^{\frac{2-\kappa}{1-\kappa}}, \text{ pour } \kappa \neq 1, 2,$$

avec les cas particulier $b(\theta) = \exp[\theta]$ pour $\kappa = 1$ (on a alors une loi quasi-Poisson) et $b(\theta) = -\log[-\theta]$ pour $\kappa = 2$ (on a alors une loi Gamma).

Si l'on suppose que une forme multiplicative $A_i \cdot B_j$ pour modéliser les incréments de paiements, on aura alors

$$\theta_{i,j} = \log[A_i \cdot B_j] \text{ si } \kappa = 1 \text{ et } \theta_{i,j} = \frac{[A_i \cdot B_j]^{1-\kappa}}{1 - \kappa} \text{ sinon.}$$

Afin de juger de la pertinence de l'ajustement, on peut calculer la **log-vraisemblance profilée** du modèle, en gardant un lien logarithmique par exemple (ce qui est parfois plus simple au niveau numérique, mais aussi au niveau de l'interprétation), ou le critère d'Akaike.

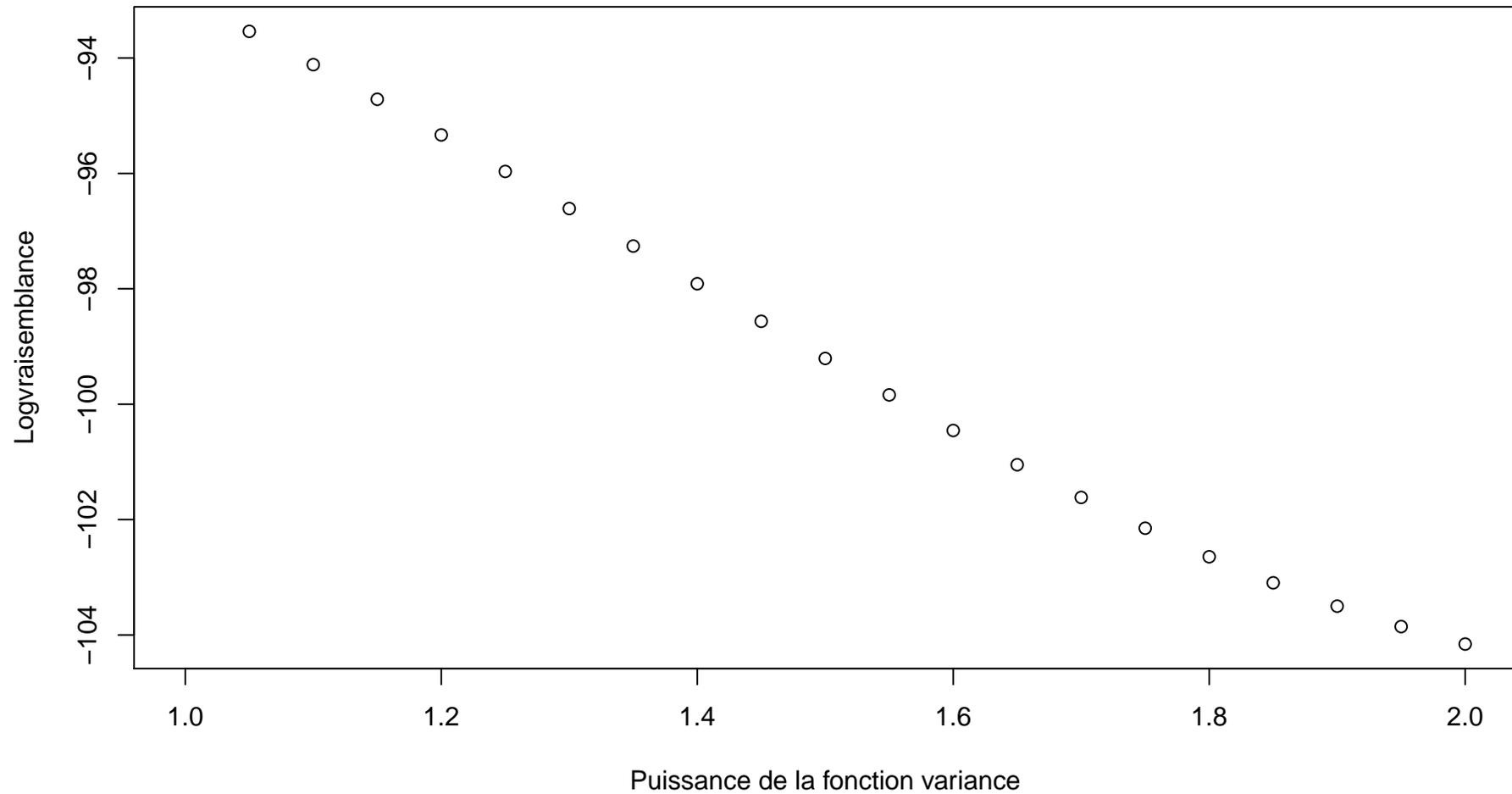


FIGURE 7 – Évolution de la log-vraisemblance profilée en fonction de κ .

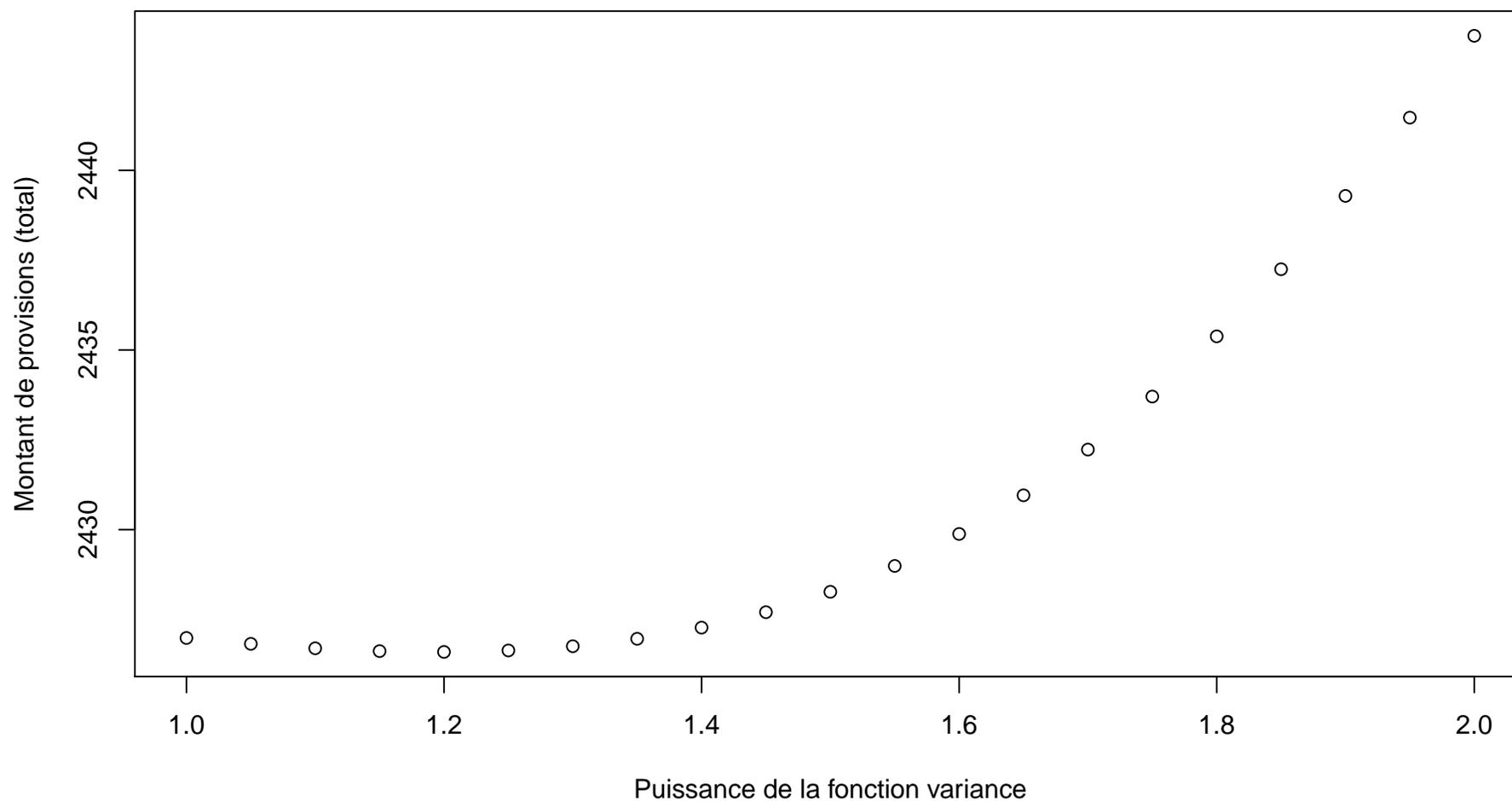


FIGURE 8 – Évolution du montant total de provision \hat{R} , en fonction de κ .

Exemple8

Si l'on souhaite garder un lien logarithmique, le paramètre le plus vraisemblable pour la fonction variance compris entre 1 et 2, est $\hat{\kappa} = 1$, ce qui correspond à un modèle de Poisson (ou plutôt quasi-Poisson).

Remarque6

Ces méthodes de simulations sont suffisamment souples pour pouvoir être adaptées, en particulier pour quantifier l'incertitude à un an. Toutefois, cette approche repose sur l'utilisation de régression GLM (Poisson ou Gamma), ce qui n'est possible que si les $Y_{i,j}$ sont positifs. En théorie, il s'agit d'incrément de paiements, qui doivent être positifs, mais en pratique, il n'est pas rare que de **incrément négatifs** soient observés. C'est d'ailleurs possible de générer un pseudo triangle possédant des incréments négatifs (e.g. [6] ou [14]).

5 Modélisation et prédiction de la mortalité future

De même que le provisionnement posait le problème de la dynamique de la vie des sinistres (dont le montant n'est pas connu le jour de la survenance du sinistre), les contrats d'assurance vie sont liés à des probabilités de décès (ou de survie) dans un futur plus ou moins lointain.

5.1 Modélisation statique des contrats d'assurance vie

En assurance vie, nous allons ici calculer des grandeurs de la forme

$$\mathbb{E} \left(\sum_{k=1}^{\infty} \frac{C_k}{(1+i)^{T_k}} \mathbf{1}(\text{paiement à la date } T_k) \right)$$

où l'assureur s'est engagé à verser un capital C_i à des dates T_i , à condition qu'une hypothèse soit vérifiée à la date T_i . Compte tenu de la linéarité de l'espérance, si l'on suppose le taux d'actualisation non aléatoire, on peut réécrire cette dernière expression sous la forme :

$$\sum_{k=1}^{\infty} \frac{C_k}{(1+i)^{T_k}} \mathbb{P}(\text{paiement à la date } T_i).$$

La valeur actuelle probable s'écrit, de manière très générale,

$$\sum_{j=1}^k \frac{C_j p_j}{(1+i)^j}$$

où $\mathbf{C} = (C_1, \dots, C_k)$ est l'ensemble des montants à verser, i est le taux d'actualisation, et $\mathbf{p} = (p_1, \dots, p_k)$ est le vecteur des probabilités de verser le capital aux différentes dates.

Exemple 9

Le plus simple est probablement la valeur actuelle probable d'un capital différé (*pure endowment*) ${}_kE_x$, correspondant à la valeur actuelle probable d'un capital de 1 dans le cas où une personne actuellement d'âge x soit encore en vie à au bout de k années, i.e.

$${}_kE_x = \frac{1}{(1+i)^k} \mathbb{P}(T > x+k | T > x) = \frac{1}{(1+i)^k} {}_kp_x$$

où ${}_kp_x$ désigne la probabilité qu'un individu d'âge x atteigne l'âge $x+k$.

Exemple 10

Considérons le cas du versement d'une unité monétaire, commençant dès aujourd'hui, et continuant tant que l'assuré sera vivant (on parlera d'annuité vie entière). On supposera l'annuité payable d'avance. On peut montrer que

$$\ddot{a}_x = \sum_{k=0}^{\infty} \frac{1}{(1+i)^k} {}^k p_x = \sum_{k=0}^{\infty} {}^k E_x$$

Plus généralement, on veut considérer non pas des assurance annuelles, mais temporaires, d'une durée de n années, i.e.

$${}_n \ddot{a}_x = \sum_{k=0}^{n-1} \frac{1}{(1+i)^k} {}^k p_x = \sum_{k=0}^{n-1} {}^k E_x$$

Notons que l'on peut également différer de h années,

$${}_{h|n} \ddot{a}_x = \sum_{k=h}^{h+n-1} \frac{1}{(1+i)^k} {}^k p_x = \sum_{k=h}^{h+n-1} {}^k E_x$$

Une table de mortalité est souvent présentée comme le suivi d'une cohorte dans le temps, avec L_0 individus à la naissance. L_k est le nombre de survivants - au sein de la cohorte - à l'âge k .

Classiquement, ${}_k p_x = \frac{L_{x+k}}{L_x}$ alors que ${}_k q_x = \frac{L_x - L_{x+k}}{L_x}$.

Par la suite, on utilisera le taux de hasard,

$${}_k p_x = \mathbb{P}_t(T > x) = \exp \left[\int_0^x \mu_y dy \right]$$

ou, si on ne considère que des valeurs discrètes

$${}_k p_x = \mathbb{P}_t(T > x) = \exp \left[\sum_{y=0}^x \mu_y \right] \text{ où } \mu_y = \log \frac{D_y}{E_y}.$$

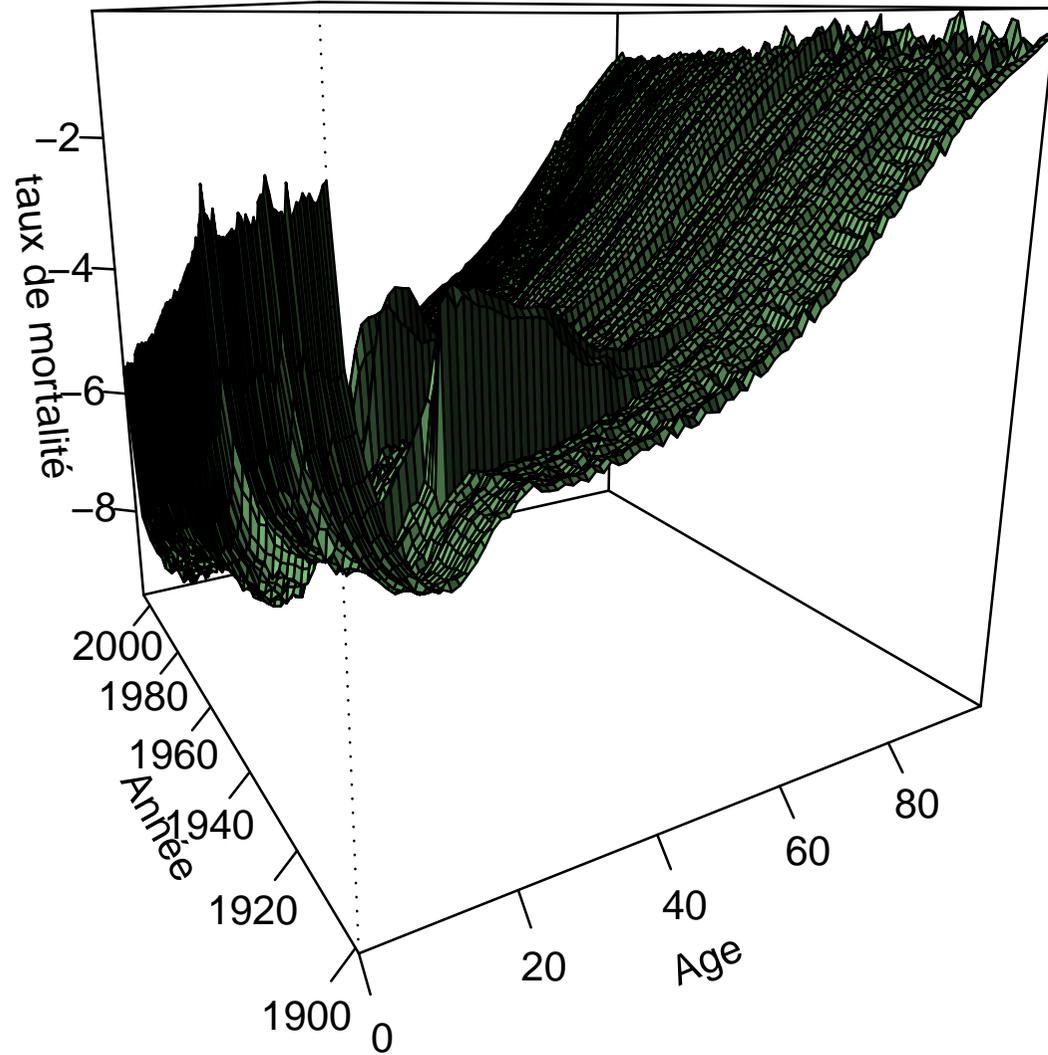
5.2 Extention dans un cadre dynamique

Dans le cadre statique, toutes les grandeurs pouvaient être construites à partir des L_x , ou des ${}_1p_x$, où x était l'âge des individus (au moment où ils souscrivaient une police d'assurance). Ici, nous allons intégrer la dimension temporelle, en notant qu'une *table* de mortalité est construite à une date t . Aussi, formellement, on notera $L_{x,t}$ le nombre de personnes d'âge x en vie à la date t .

Nous disposons ici de données françaises, avec la mortalité des femmes, des hommes, et de l'ensemble, entre 1899 et 2005. Ici on dispose de $D_{x,t}$ le nombre de personnes décédée à l'âge x l'année t , et $E_{x,t}$ l'exposition.

Pour commencer, on peut visualiser l'évolution de la surface du taux de mortalité, afin de mieux comprendre la nécessité d'une analyse dynamique de la démographie, où

$$\mu_{x,t} = \log \left(\frac{D_{x,t}}{E_{x,t}} \right)$$



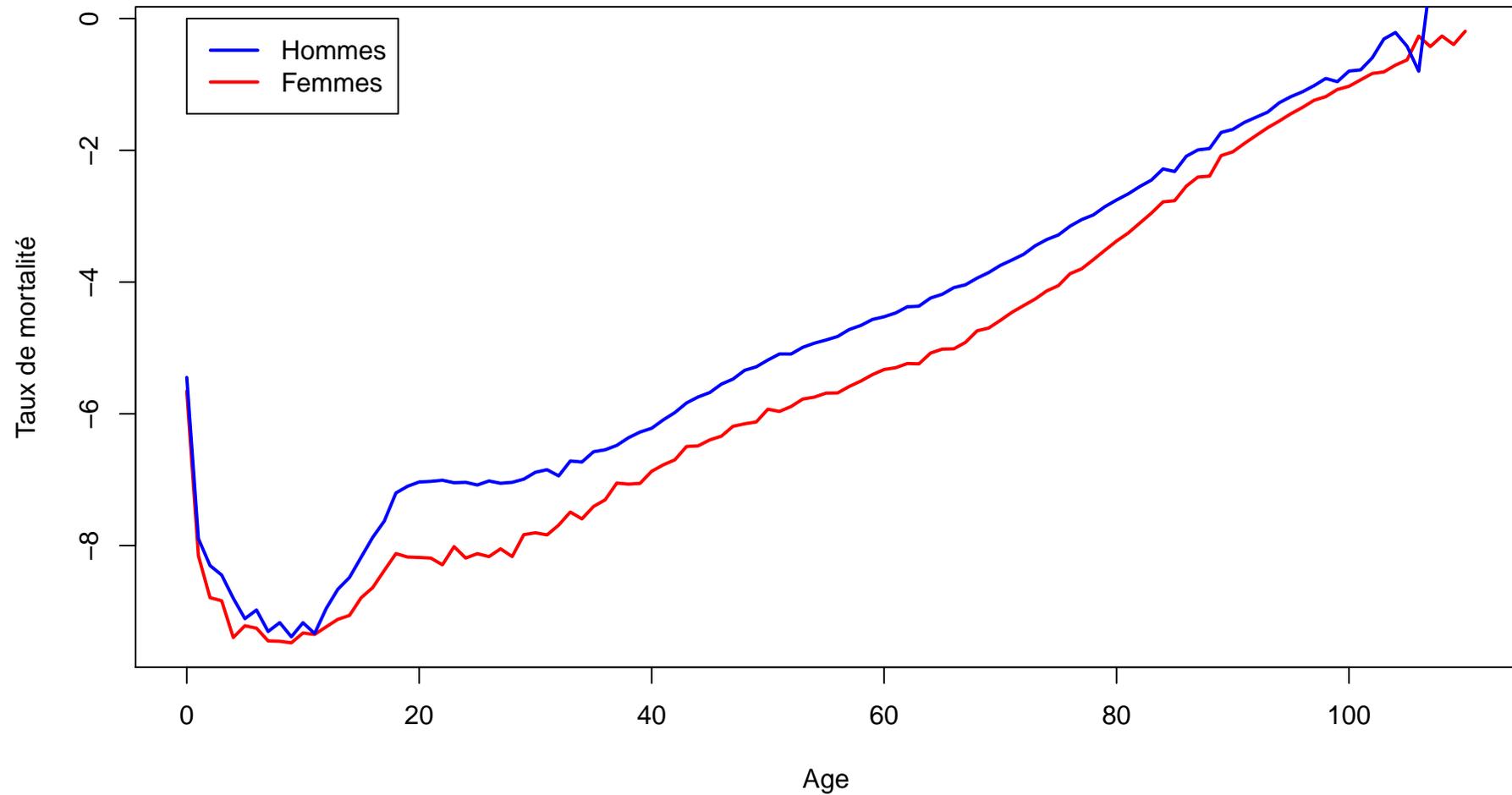


FIGURE 9 – Taux de hasard $x \mapsto \mu_{x,t}$ en $t = 2000$.

Compte tenu du lien entre le taux de hasard et les fonctions de survie, on peut en déduire les fonctions de survie à la naissance (c'est à dire $x = 0$),

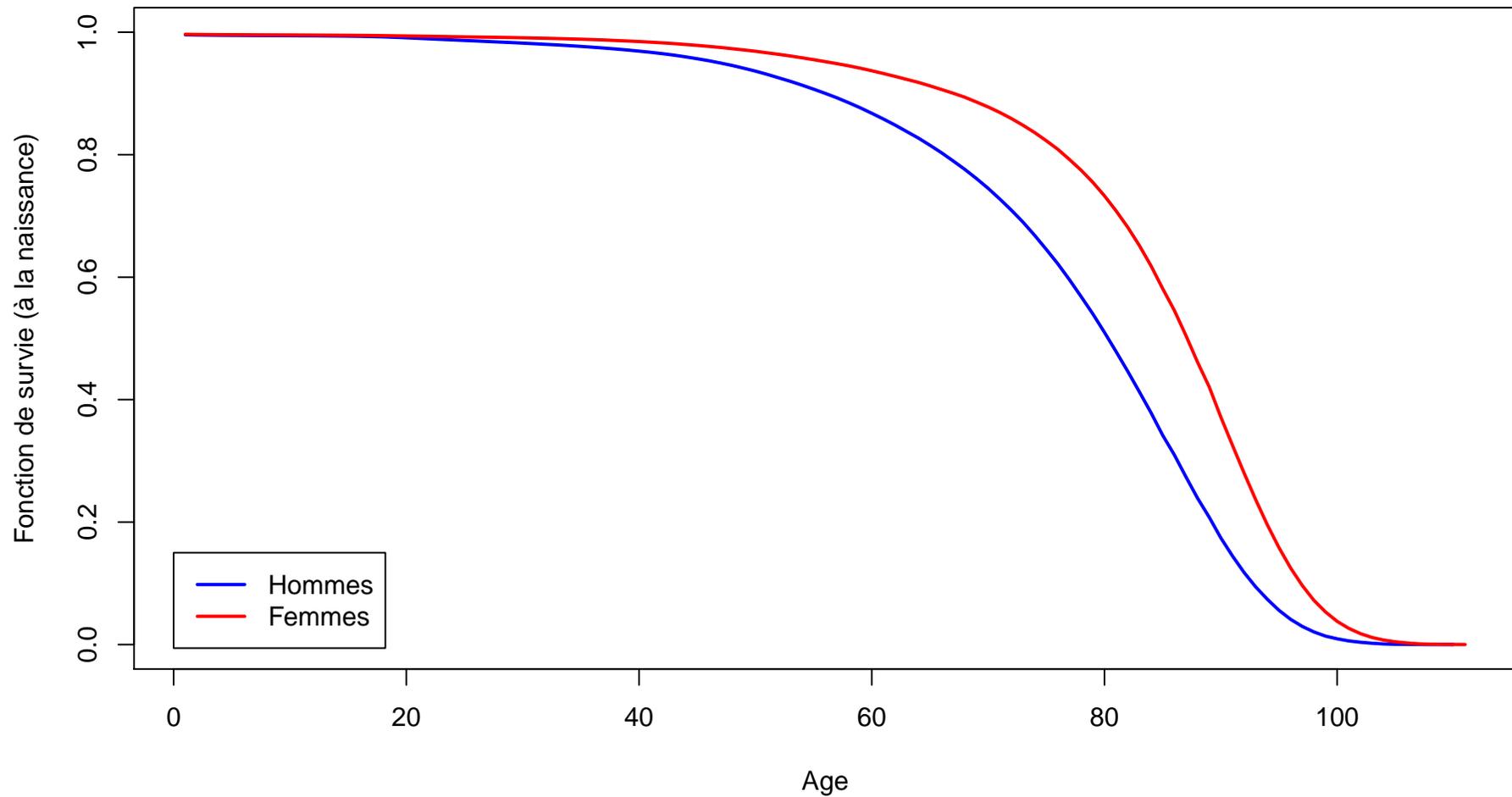


FIGURE 10 – Survie (à la naissance), en $t = 2000$, i.e. $x \mapsto \mathbb{P}_t(T > x)$.

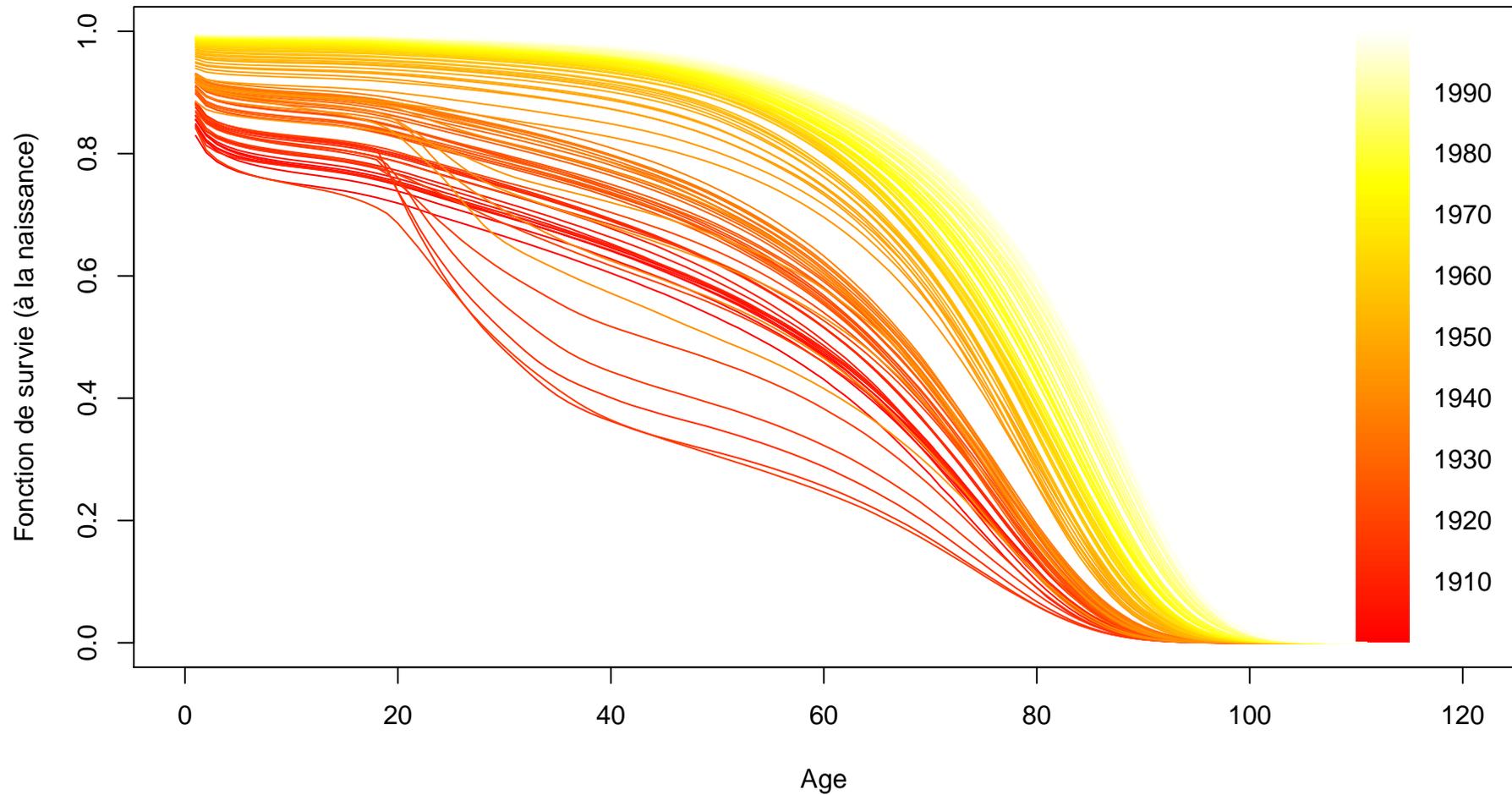


FIGURE 11 – Rectangularisation des fonctions de survie à la naissance.

5.3 La lecture transversale des tables

En fait, cette lecture transversale des tables - bien que correspondant à ce que nous avons fait jusqu'à présent - ne paraît pas valide.

Aussi, afin de lire la fonction de survie pour un individu (ou une cohorte), on ne lit plus la base par année, mais suivant une diagonale.

6 Le modèle de Lee & Carter

La modélisation retenue pour le taux instantané de mortalité est

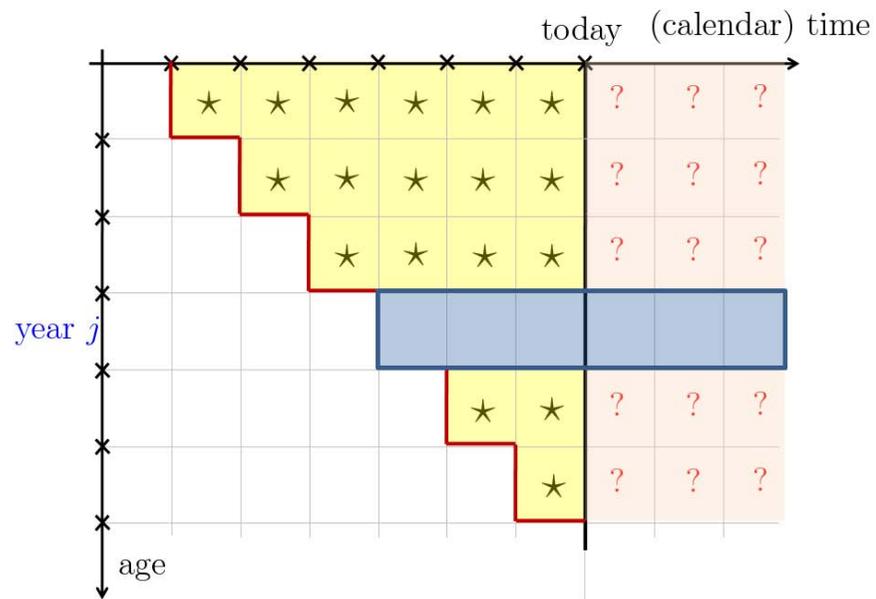
$$\log \mu_{x,t} = \alpha_x + \beta_x \kappa_t + \varepsilon_{x,t},$$

avec les variables aléatoires $\varepsilon_{x,t}$ i.i.d.; l'idée du modèle est donc d'ajuster à la série (doublement indicée par x et t) des logarithmes des taux instantanés de décès une structure paramétrique (déterministe) à laquelle s'ajoute un phénomène aléatoire; le critère d'optimisation retenu va consister à maximiser la variance expliquée par le modèle, ce qui revient à minimiser la variance des erreurs.

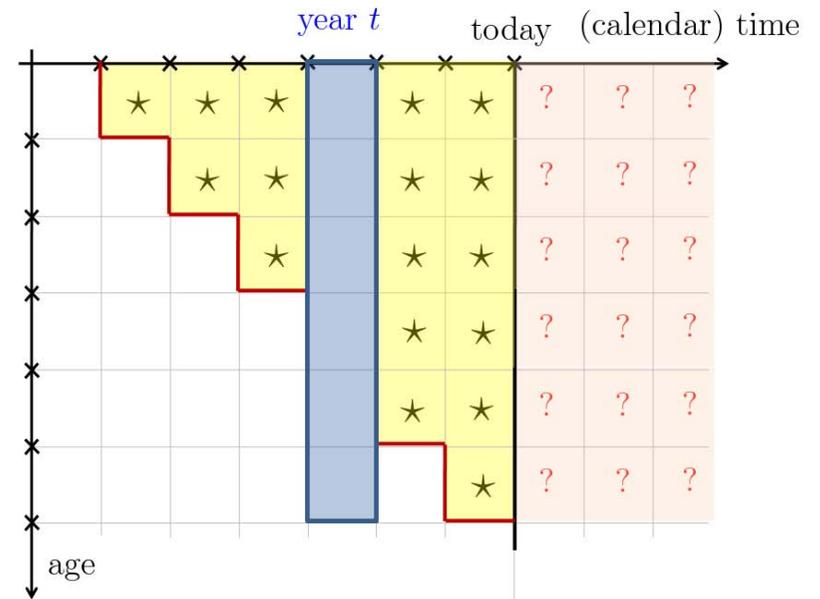
On retient en général les deux contraintes d'identifiabilité suivantes :

$$\sum_{x=0}^{120} \beta_x = 1 \text{ et } \sum_{t=1900}^{2005} \kappa_t = 0.$$

the **age** factors (α_x, β_x)

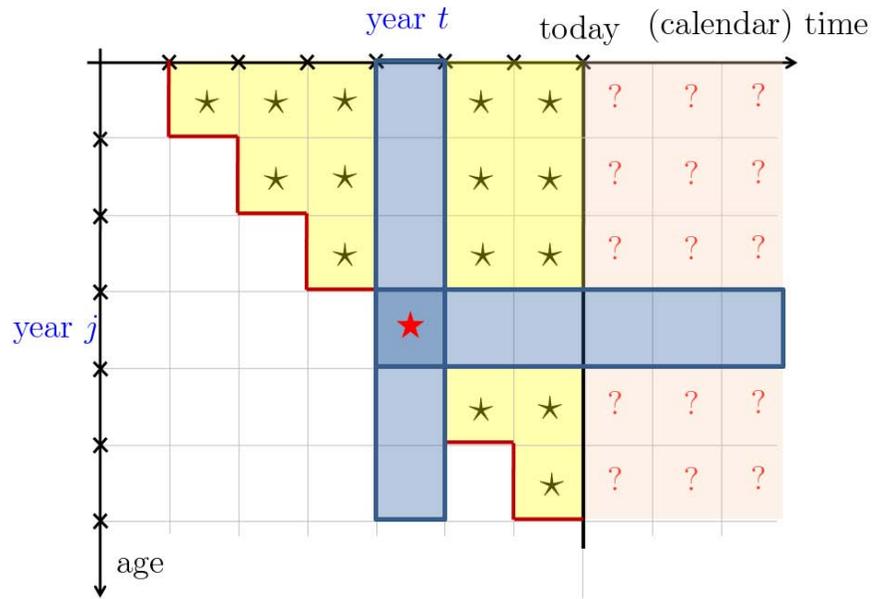


the **time** factor t



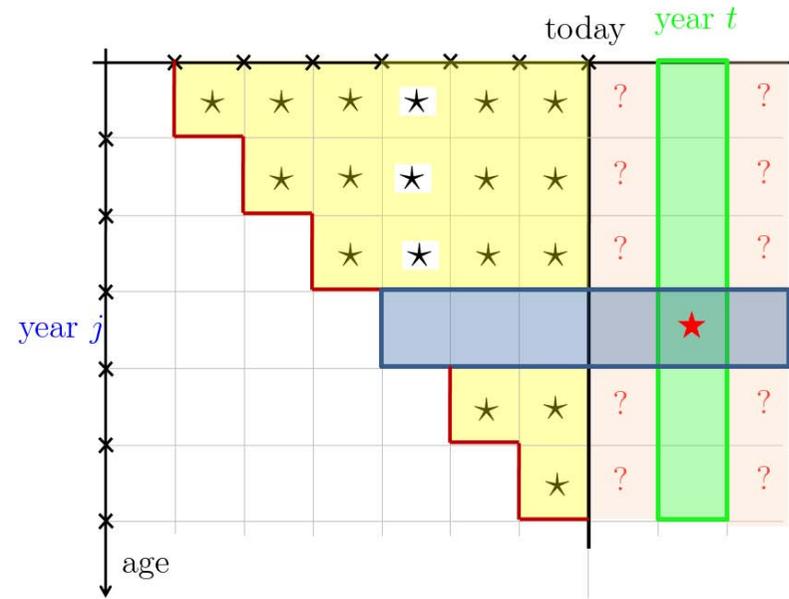
$$\exp[\hat{\alpha}_j + \hat{\beta}_j \hat{\gamma}_t]$$

on past observations



$$\exp[\hat{\alpha}_j + \hat{\beta}_j \tilde{\gamma}_t]$$

on the future



L'estimation des paramètres s'effectue en résolvant (numériquement)

$$\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\kappa}}\right) = \arg \min \sum_{x,t} (\log \mu_{xt} - \alpha_x - \beta_x \kappa_t)^2,$$

où $\boldsymbol{\alpha} = (\alpha_x)$, $\boldsymbol{\beta} = (\beta_x)$ et $\boldsymbol{\kappa} = (\kappa_t)$. Ces coefficients sont représentés sur la Figure 12.

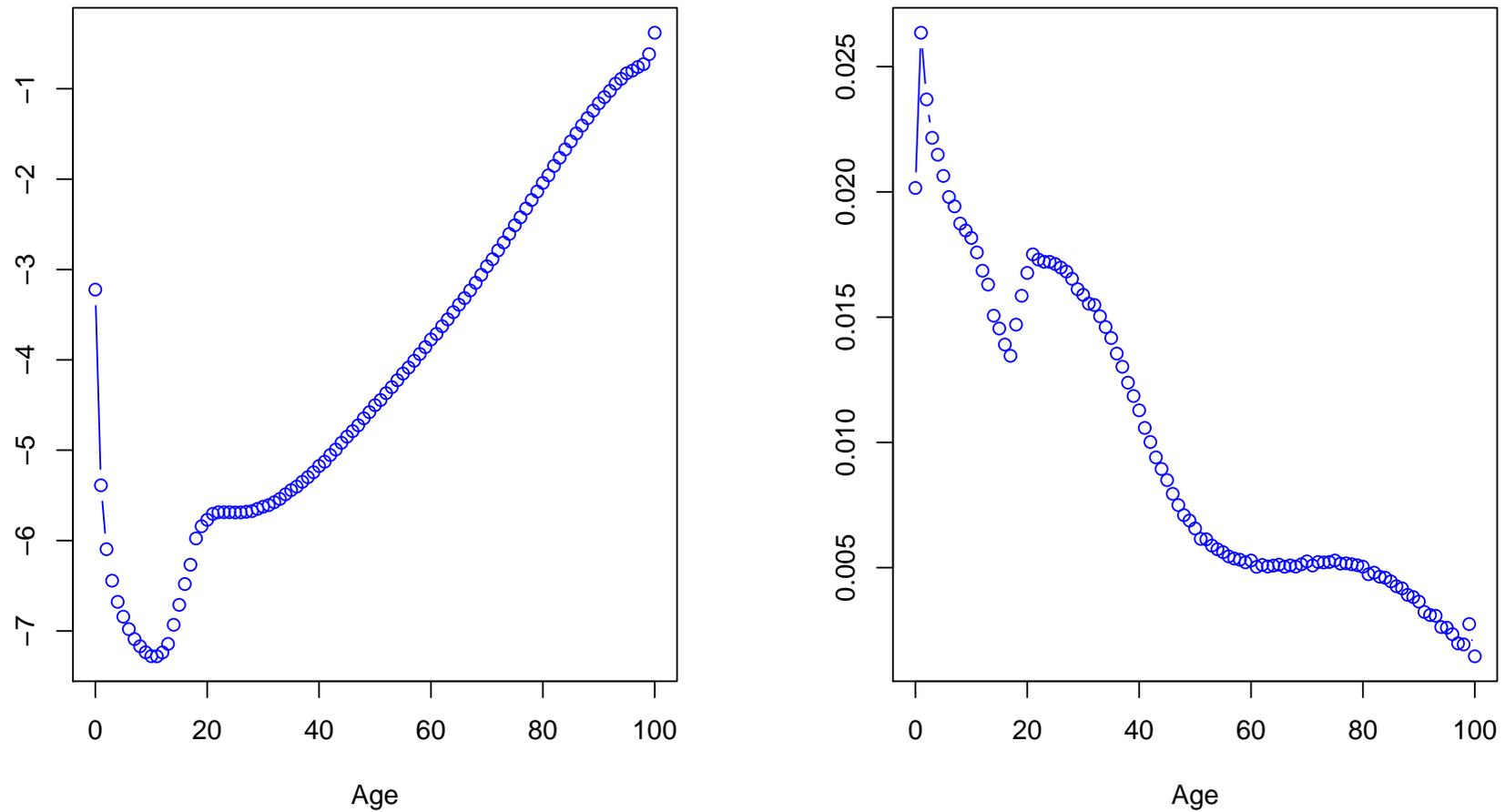


FIGURE 12 – Estimation des coefficients $\hat{\alpha} = (\hat{\alpha}_x)$ et $\hat{\beta} = (\hat{\beta}_x)$.

Une fois l'ajustement réalisé sur les données disponibles, on peut réaliser des projections de la mortalité future, c'est à dire du coefficient κ_t , comme sur la Figure 13

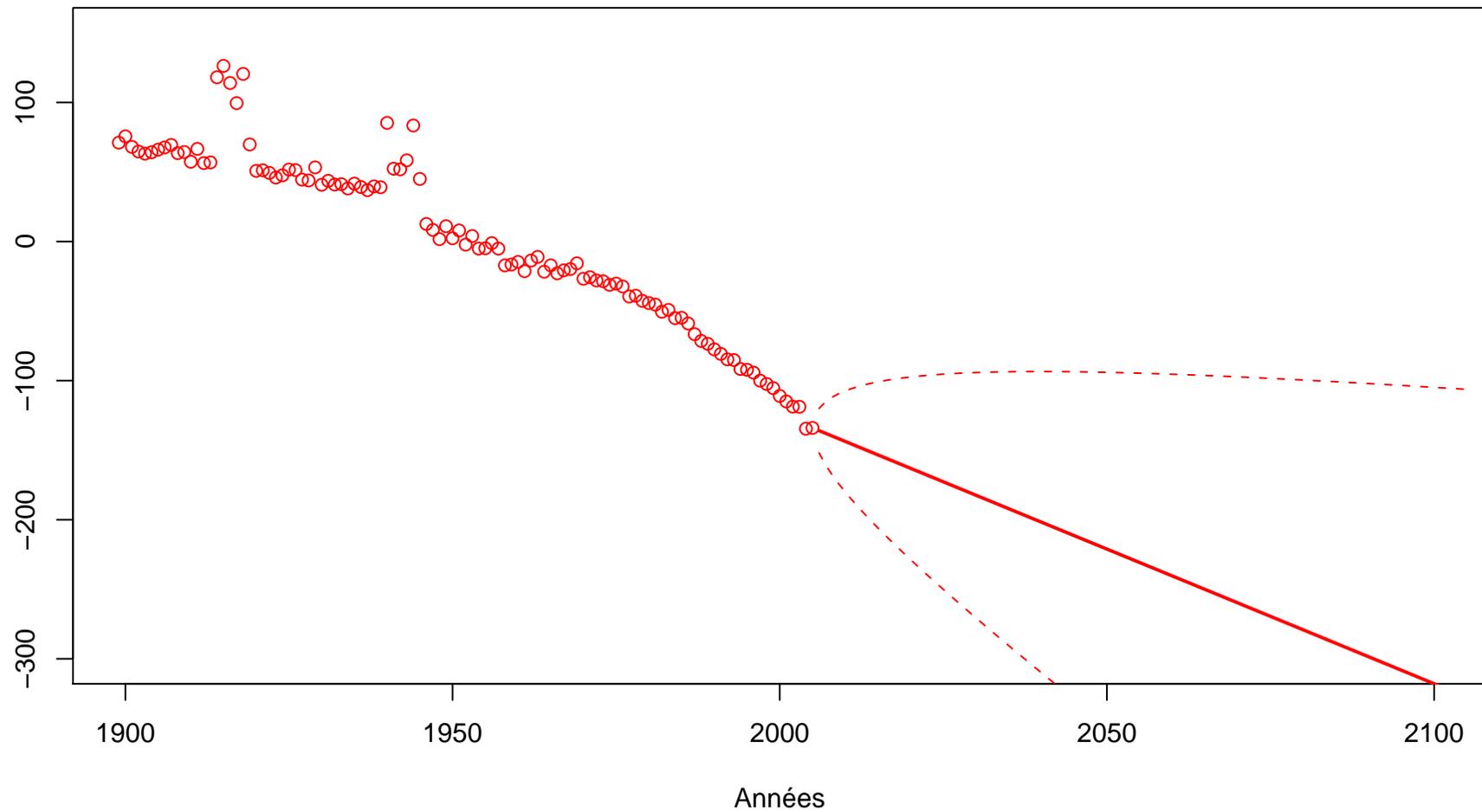


FIGURE 13 – Estimation et projection des coefficients $\hat{\kappa} = (\hat{\kappa}_t)$.

On peut utiliser ces prédictions (en l'occurrence à l'aide d'un processus $ARIMA(1, 1, 0)$ pour prédire l'évolution l'espérance de vie des assurés,

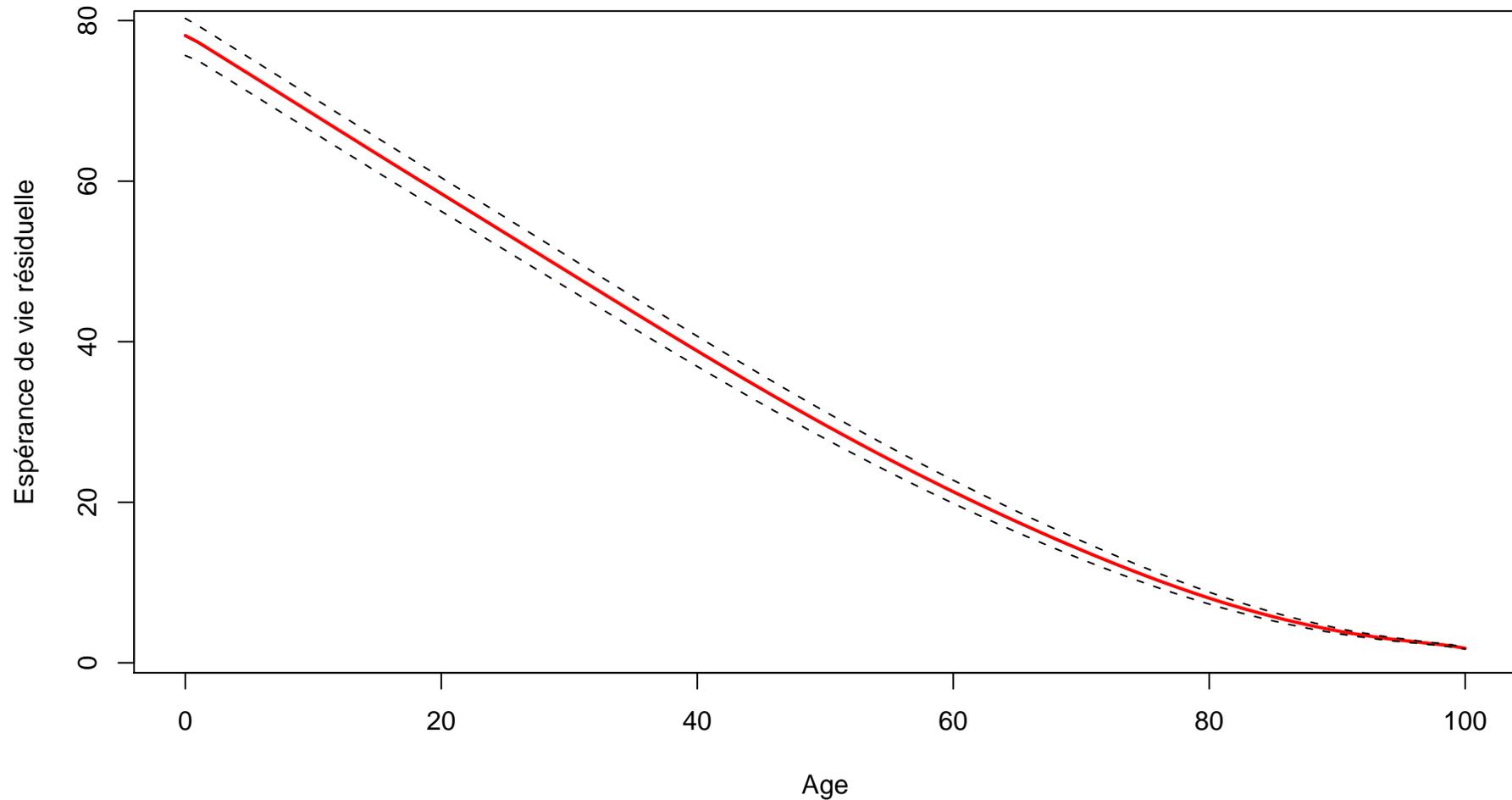


FIGURE 14 – Espérances de vie (résiduelle) à l'âge x , à la date $t = 2000$.

Mais au delà des travaux sur les $\mu_{x,t}$, on peut également travailler sur une modélisation des *nombres* de décès $D_{x,t}$. Et le modèle de Lee & Carter (qui repose sur une idée de modèle Gaussien pour $\mu_{x,t}$ peut alors servir de base.

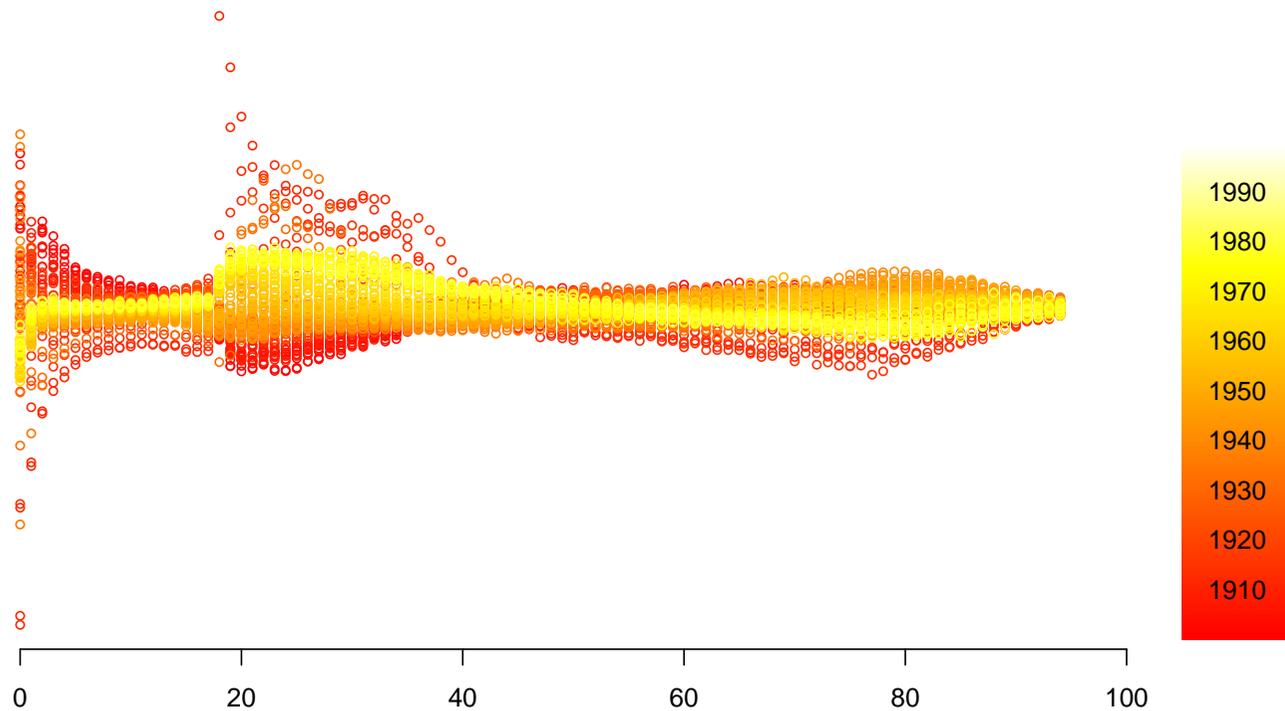
L'approche proposée par [3], [9], [5] ou encore [21] consiste à supposer (comme sur les modèles de provisions) que l'on peut modéliser le nombre de décès à l'aide d'un modèle GLM Poissonien. Or compte tenu de l'analyse précédente sur les taux de décès $\mu_{x,t}$, il est légitime de considérer un modèle

$$D_{x,t} \sim \mathcal{P}(E_{x,t} \exp[\mu_{x,t}]).$$

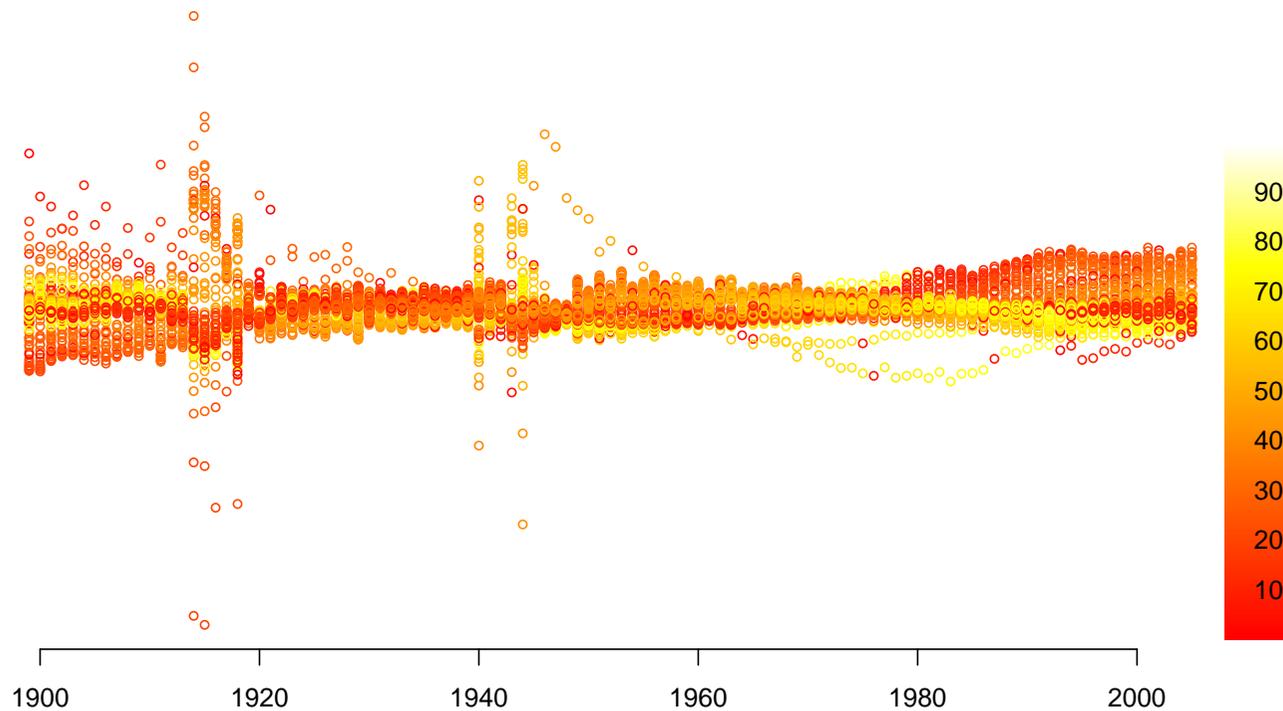
On considère alors une régression Poissonnienne, avec le logarithme de l'exposition comme variable offset, mais compte tenu du modèle de Lee & Carter, il ne s'agit plus d'un modèle GLM, les facteurs α , β et κ n'intervenant plus de manière linéaire. Mais il est toujours possible d'estimer les paramètres à l'aide d'algorithmes numériques.

Et de même que pour les modèles de provisionnement on peut alors commencer une analyse des résidus,

Résidus de Pearson , $\varepsilon_{x,t} = \frac{D_{x,t} - \hat{D}_{x,t}}{\sqrt{\hat{D}_{x,t}}}$, en fonction de l'âge x



Résidus de Pearson , $\varepsilon_{j,t} = \frac{D_{j,t} - \hat{D}_{j,t}}{\sqrt{\hat{D}_{j,t}}}$, en fonction du temps t



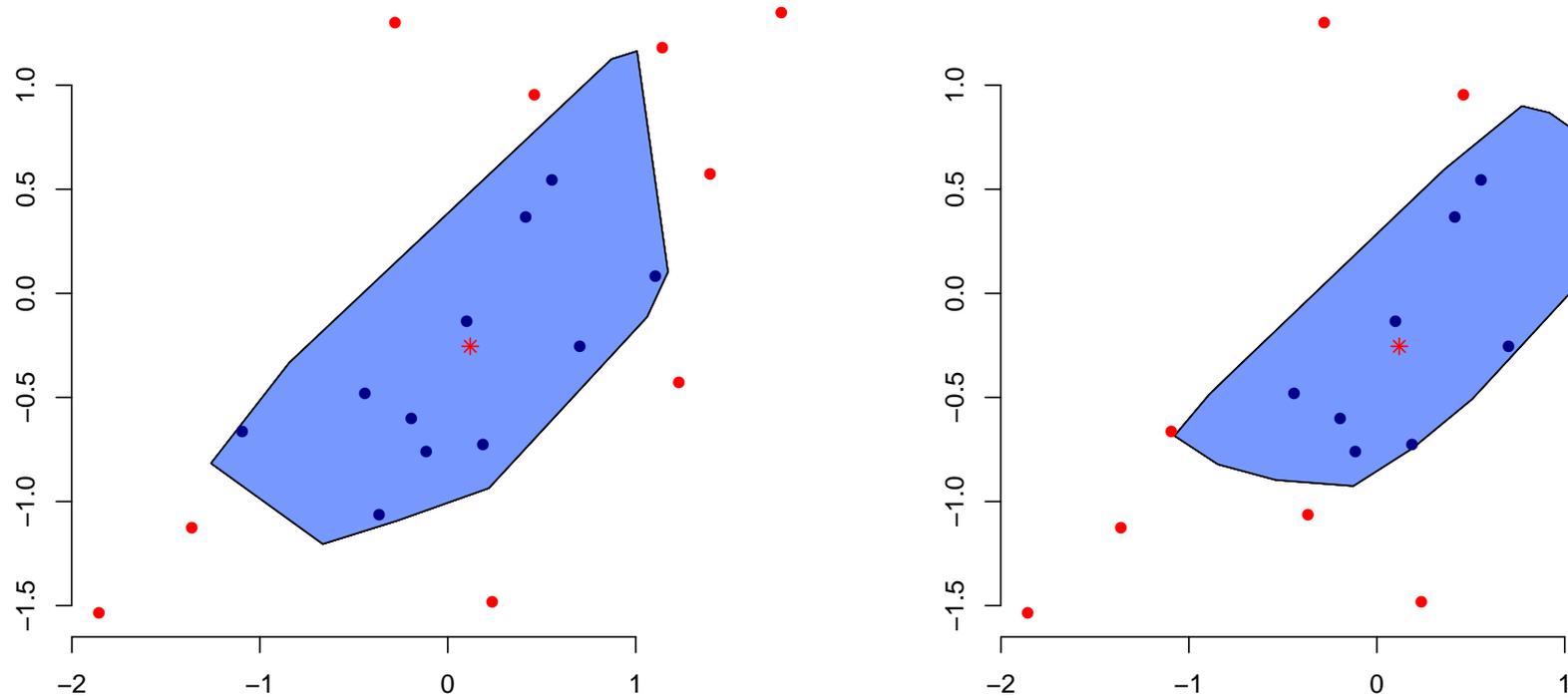
Outliers can *simply* be understood in a univariate context. To extend it in higher dimension, [27] defined the

$$\text{depth}(\mathbf{y}) = \min_{\mathbf{u}, \mathbf{u} \neq \mathbf{0}} \left\{ \frac{1}{n} \sum_{i=1}^n \mathbf{1}(\mathbf{X}_i \in H_{\mathbf{y}, \mathbf{u}}) \right\}$$

where $H_{\mathbf{y}, \mathbf{u}} = \{\mathbf{x} \in \mathbb{R}^d \text{ such that } \mathbf{u}'\mathbf{x} \leq \mathbf{u}'\mathbf{y}\}$ and for $\alpha > 0.5$, defined the **depth set** as

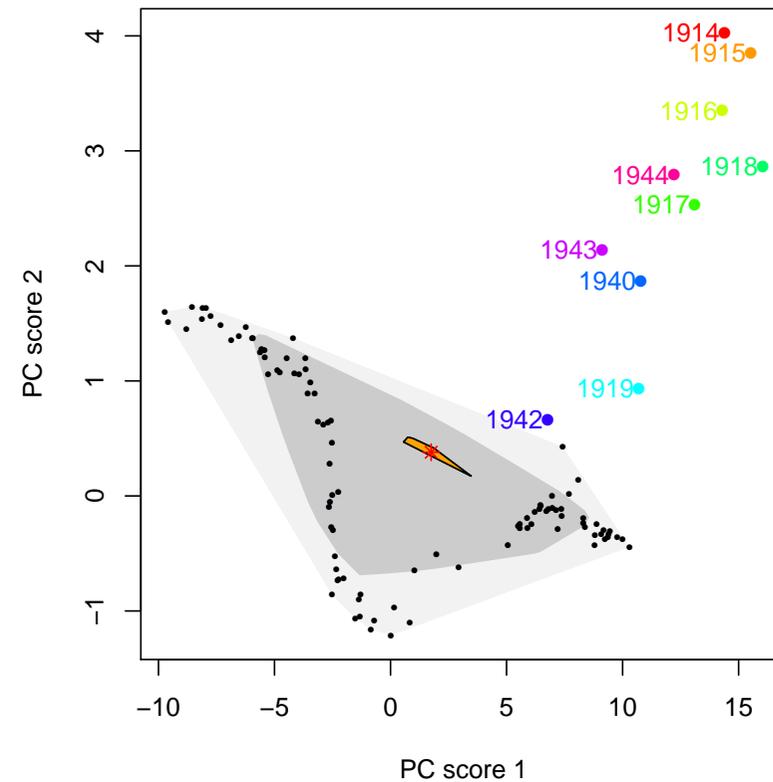
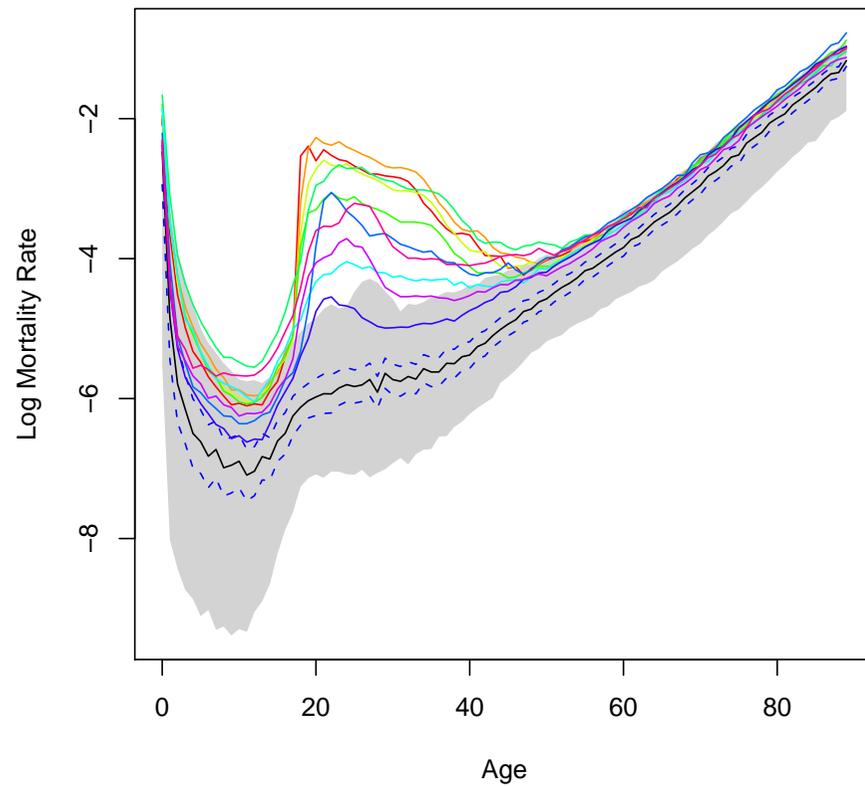
$$D_\alpha = \{\mathbf{y} \in \mathbb{R} \in \mathbb{R}^d \text{ such that } \text{depth}(\mathbf{y}) \geq 1 - \alpha\}.$$

The empirical version is called the **bagplot** function (see e.g. [22]).



where the blue set is the empirical estimation for D_α , $\alpha = 0.5$.

It is possible to extend it to define (past) **functional outliers**,



(here male log-mortality rates in France from 1899 to 2005).

- [1] E. Astesan. *Les réserves techniques des sociétés d'assurance contre les accidents automobiles*. Librarite Générale de Droit & de Jurisprudence, Paris, 1938.
- [2] R.A. Bailey. Insurance rates with minimum bias. *Proceedings of the Society of Actuaries*, 50 :4–11, 1963.
- [3] Natacha Brouhns, Michel Denuit, and Jeroen K. Vermunt. A poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance : Mathematics and Economics*, 31(3) :373 – 393, 2002.
- [4] S. Christofides. Regression models based on log-incremental payments. In Institute of Actuaries, editor, *Claims Reserving Manual*, 1989.
- [5] Iain Currie, James Kirkby, Maria Durban, and Paul Eilers. Smooth lee-carter models and beyond. In *Annual meeting*. Society of Actuaries, 2004.
- [6] Enrique de Alba. Claims reserving when there are negative values in the runoff triangle : Bayesian analysis using the three-parameter log-normal distribution. *N. Am. Actuar. J.*, 10(3) :45–59, 2006.

- [7] M. Denuit and A. Charpentier. *Mathématiques de l'assurance non-vie : Tarification et provisionnement. Tome 2*. Economica, 2005.
- [8] P. D. England and R. J. Verrall. Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance : Mathematics and Economics*, 25 :281–293, 1999.
- [9] S. Haberman and A. Renshaw. Mortality, longevity and experiments with the lee-carter model. *Lifetime Data Analysis*, 2008.
- [10] C. A. Hachemeister and J. N. Stanard. Ibrn claims count estimation with static lag functions. In *12th ASTIN Colloquium*, Portimao, Portugal, 1975.
- [11] Henry. 1937.
- [12] B. JØorgensen. *The theory of dispersion models*. Chapman & Hall, 1997.
- [13] E. Kremer. Ibrn claims and the two-way model of anova. *Scandinavian Actuarial Journal*, pages 47–55, 1982.
- [14] Michael Kunkler. Modelling negatives in stochastic reserving models. *Insurance : Mathematics and Economics*, 38(3) :540 – 555, 2006.

- [15] Ronald D. Lee, Lawrence R. Carter, Robert McNown, and Juha M. Alho. Modeling and forecasting U.S. mortality. 87(419) :659–??, September 1992.
- [16] T. Mack. A simple parametric model for rating automobile insurance or estimating ibnr claims reserves. *ASTIN Bulletin*, 21 :93–109, 1991.
- [17] T. Mack. Distribution-free calculation of the standard error of chain-ladder reserve estimates. *ASTIN Bulletin*, 15 :133–138, 1993.
- [18] T. Mack. The standard error of chain-ladder reserve estimates : Recursive calculation and inclusion of a tail factor. *ASTIN Bulletin*, 29 :361–366, 1993.
- [19] M. Merz and M. V Wüthrich. Modelling the claims development result for solvency purposes. *CAS E-Forum*, pages 542–568, 2008.
- [20] A. E. Renshaw and R. J. Verrall. A stochastic model underlying the chain-ladder technique. *British Actuarial Journal*, 4 :903–923, 1998.
- [21] A.E. Renshaw and S. Haberman. A cohort-based extension to the lee-carter model for mortality reduction factors. *Insurance : Mathematics and Economics*, 38(3) :556 – 570, 2006.

- [22] Ruts I. Tukey J.W. Rousseeuw, P.J. The bagplot : A bivariate boxplot. *Annals of Statistics*, 53 :382–387, 1999.
- [23] Klaus Schmidt and Angela Wünsche. Chain ladder, marginal sum and maximum likelihood estimation. *Blätter der DGVM*, 23 :267–277, 1998. 10.1007/BF02808289.
- [24] Klaus D. Schmidt. A note on the overdispersed poisson family. *Insurance : Mathematics and Economics*, 30(1) :21 – 25, 2002.
- [25] G. Simonet. *Comptabilité des entreprises d'assurance*. L'Argus de l'Assurance, 1998.
- [26] G. Taylor. Separation of inflation and other effects from the distribution of non-life insurance claims delays. *ASTIN Bulletin*, 9(1) :219–231, 1977.
- [27] J.W. Tukey. Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians*, 2 :523–531, 1975.