



Méthodes stochastiques de calcul des réserves en assurance non vie

Esterina Masiello

M1 Actuariat

2020/2021



Limites des méthodes déterministes

- impossibilité d'obtenir une estimation de la loi de probabilité de la provision totale;
- pas de calcul de volatilité, VaR, ...
- impossibilité de mesurer l'incertitude sur l'estimation des provisions de sinistres;
- ...

⇒ Méthodes stochastiques



Méthodes stochastiques

Idée générale : les éléments du triangle de liquidation (paiements cumulés, incrémentaux, etc.) sont considérés comme des réalisations de variable aléatoire réelles (observées à la date de fin d'inventaire : 31/12/n).

Risque : risque de modèle, i.e. le risque d'erreur de spécification (on utilise un modèle inexacte qui produit des résultats erronés).



Méthodes stochastiques

- a) Modèle recursif de Mack
- b) Modèles stochastiques factoriels
 - Régression lognormale
 - Modèle Poissonien
- c) Distribution prédictives et VaR
 - 1) Distribution prédictives par convolution
 - 2) Distribution prédictives approchées
 - 3) Distribution prédictives par bootstrap

a) Modèle recursif de Mack

Il s'agit d'un Chain Ladder stochastique !

Hypothèses du modèle :

- H1) Indépendance ligne par ligne : (C_{i1}, \dots, C_{in}) et (C_{j1}, \dots, C_{jn}) sont des vecteurs de v.a. indépendants pour $i \neq j$
- H2) Il existe une constante $f_j > 0$ telle que $\forall j = 1, \dots, n - 1$ et $\forall i = 1, \dots, n$

*ici on va de 1 à m
Or avant on résomait de 0 à m-1*

$$E(C_{i,j+1} \mid C_{i1}, \dots, C_{ij}) = f_j C_{ij}$$

Remarque: L'hypothèse H_1 peut ne pas être vérifiée en pratique en cas, par exemple, de changements importants au niveau de la gestion des sinistres ou du taux d'inflation de la branche étudiée puisque cela va avoir un impact sur plusieurs exercices d'origine.

Pour éviter ce problème, il faudrait travailler avec des **DONNEES AS IF**.

Le montant "as if" d'un sinistre est le coût de celui-ci s'il surviennent avec les mêmes caractéristiques mais dans l'environnement présent à la date de l'étude (31/12/m).

Proposition 1 Soit $T = \{C_{ij} \mid i + j \leq n + 1\}$. Sous les hypothèses H1) et H2), on obtient pour $i \geq 2$

$$E(C_{in} \mid T) = f_{n-1} \cdots f_{n-i+1} C_{i,n-i+1}$$

Preuve Par H1) on a :

$$E(C_{in} \mid T) = E(C_{in} \mid C_{i1}, \dots, C_{i,n-i+1})$$

et, par double conditionnement

$$E(C_{in} \mid T) = E(E(C_{in} \mid C_{i1}, \dots, C_{i,n-1}) \mid C_{i1}, \dots, C_{i,n-i+1})$$

mais, par H2), $E(C_{in} \mid C_{i1}, \dots, C_{i,n-1}) = f_{n-1} C_{i,n-1}$ et donc

$$E(C_{in} \mid T) = f_{n-1} E(C_{i,n-1} \mid C_{i1}, \dots, C_{i,n-i+1})$$

et, de proche en proche,

$$\begin{aligned} E(C_{in} \mid T) &= f_{n-1} \cdot \dots \cdot f_{n-i+1} E(C_{i,n-i+1} \mid C_{i1}, \dots, C_{i,n-i+1}) \\ &= f_{n-1} \cdot \dots \cdot f_{n-i+1} C_{i,n-i+1} \quad \square \end{aligned}$$

Le théorème montre que l'estimateur

$$\hat{C}_{in} = C_{i,n-i+1} \hat{f}_{n-i+1} \cdots \hat{f}_{n-1}$$

a la même forme que $E(C_{in} \mid T)$ qui est la meilleure prédition de C_{in} basée sur les observations T .

Les estimateurs des facteurs de développement sont donnés par le modèle CL :

$$\hat{f}_j = \frac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{ij}} \quad j = 1, \dots, n-1$$

Proposition 2 Sous les hypothèses H1) et H2), les estimateurs des facteurs de développement \hat{f}_j sont sans biais et non corrélés.

Preuve On note, pour $j = 1, \dots, n$

$$T_j = \{C_{ih} \mid h \leq j, i + j \leq n + 1\}$$

D'après H2), on a :

$$\begin{aligned} E(C_{i,j+1} \mid T_j) &= E(C_{i,j+1} \mid C_{i1}, \dots, C_{ij}) \\ &= f_j C_{ij} \end{aligned}$$

$$\begin{aligned}
 E(\hat{f}_j \mid T_j) &= E\left[\frac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{ij}} \mid T_j\right] \\
 &= \frac{\sum_{i=1}^{n-j} E(C_{i,j+1} \mid T_j)}{\sum_{i=1}^{n-j} C_{ij}} \\
 &= f_j
 \end{aligned}$$

et, par conditionnement,

$$E(\hat{f}_j) = E(E(\hat{f}_j \mid T_j)) = f_j$$

\hat{f}_j est donc un estimateur sans biais de f_j .

De plus, pour $j < k$,

$$\begin{aligned} E(\hat{f}_j \hat{f}_k) &= E(E(\hat{f}_j \hat{f}_k \mid T_k)) \\ &= E[\hat{f}_j E(\hat{f}_k \mid T_k)] \\ &= E(\hat{f}_j) f_k \\ &= E(\hat{f}_j) E(\hat{f}_k) \quad \square \end{aligned}$$

Cette absence de corrélation entre les estimateurs est centrale puisqu'elle permet d'écrire :

$$E(\hat{f}_j \hat{f}_{j+1} \cdots \hat{f}_{k-1} \hat{f}_k) = f_j f_{j+1} \cdots f_{k-1} f_k$$

ce qui permet de transmettre l'absence de biais aux estimateurs calculés à l'aide des facteurs chain ladder. Par exemple

$$\hat{C}_{in} = C_{i,n-i+1} \hat{f}_{n-i+1} \cdots \hat{f}_{n-1}$$

est un estimateur sans biais de

$$E(C_{in} \mid T) = C_{i,n-i+1} f_{n-i+1} \cdots f_{n-1}$$

$\hat{R}_i = \hat{C}_{im} - \hat{C}_{i,m-i+1}$ est un estimateur sans biais de R_i car $E[\hat{R}_i] = R_i$

Erreur de prévision

Classiquement pour quantifier l'erreur associée à un estimateur $\hat{\theta}$ d'un paramètre θ , on calcule le MSE associé

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

Or, ici on cherche à quantifier l'incertitude de l'estimateur \hat{R}_i associé à une variable aléatoire R_i . Le MSE s'écrit alors :

$$MSE(\hat{R}_i) = E[(\hat{R}_i - E(R_i))^2]$$

Si l'on souhaite comparer à R_i , on ne parle pas de MSE mais de MSEP (erreur quadratique moyenne de prédiction)

$$MSEP(\hat{R}_i) = E[(\hat{R}_i - R_i)^2]$$

Erreur de prévision

L'erreur quadratique moyenne de prédiction conditionnelle (MSECP) est définie par

$$MSECP(\hat{R}_i) = E[(\hat{R}_i - R_i)^2 \mid T_i]$$

Il est possible de montrer que :

$$E[(\hat{R}_i - R_i)^2] \approx E[(\hat{R}_i - E(R_i))^2] + E[(R_i - E(R_i))^2]$$

avec

$$E[(\hat{R}_i - E(R_i))^2] = MSE(\hat{R}_i)$$

l'erreur d'estimation et

$$E[(R_i - E(R_i))^2] = Var(R_i)$$

l'erreur classique du modèle.



Hypothèse supplémentaire

H3) pour $j = 1, \dots, n - 1$, il existe une constante σ_j^2 telle que

$$V(C_{i,j+1} \mid C_{i1}, \dots, C_{ij}) = \sigma_j^2 C_{ij} \quad i = 1, \dots, n$$

Proposition 3 Sous les hypothèses H1), H2) et H3),

$$\widehat{MSEP}(\hat{R}_i) = \hat{C}_{in}^2 \sum_{j=n-i+1}^{n-1} \frac{\hat{\sigma}_j^2}{\hat{f}_j^2} \left[\frac{1}{\hat{C}_{ij}} + \frac{1}{\sum_{k=1}^{n-j} C_{kj}} \right]$$

avec la convention $\hat{C}_{i,n-i+1} = C_{i,n-i+1}$ et

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

pour $j \leq n-2$. Pour $j = n-1$,

$$\hat{\sigma}_j^2 = \min \left\{ \frac{\hat{\sigma}_{n-2}^4}{\hat{\sigma}_{n-3}^2}, \min \left\{ \hat{\sigma}_{n-3}^2, \hat{\sigma}_{n-2}^2 \right\} \right\}$$

La valeur de $\hat{\sigma}_j^2$ pour $j = n - 1$ est extrapolée à partir de la série $\hat{\sigma}_1^2, \dots, \hat{\sigma}_{n-3}^2, \hat{\sigma}_{n-2}^2$ de telle sorte que

$$\frac{\hat{\sigma}_{n-3}^2}{\hat{\sigma}_{n-2}^2} = \frac{\hat{\sigma}_{n-2}^2}{\hat{\sigma}_{n-1}^2}$$

est vérifié pour $\hat{\sigma}_{n-3} > \hat{\sigma}_{n-2}$.

Preuve : voir T. Mack (1993) *Distribution-free calculation of the standard error of chain-ladder reserve estimates*

Réécriture du modèle

$$C_{i,j+1} = f_j C_{ij} + \sigma_j^2 \sqrt{C_{ij}} \epsilon_{ij}$$

avec ϵ_{ij} i.i.d. centrés et de variance unitaire. On estime les paramètres inconnus par la méthode des MCP, i.e. en cherchant à résoudre :

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

Les résidus standardisés

$$\hat{\epsilon}_{ij} = \frac{C_{i,j+1} - \hat{f}_j C_{ij}}{\sqrt{C_{ij}}}$$

nous donnent une idée simple pour estimer le paramètre de volatilité

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

Corollaire Sous les mêmes hypothèses que la Proposition 3, on obtient

$$\widehat{MSEP}(\hat{R}) = \sum_{i=2}^n \widehat{MSEP}(\hat{R}_i) + \hat{C}_{in} \left(\sum_{k=i+1}^n \hat{C}_{kn} \right) \sum_{j=n-i+1}^{n-1} \frac{2\hat{\sigma}_j^2}{\hat{f}_j^2 \sum_{h=1}^{n-j} C_{hj}}$$

Estimation des quantiles

Le modèle de Mack ne permet pas d'obtenir la description complète de la distribution des provisions mais en faisant une hypothèse sur la distribution des provisions, on peut obtenir les quantiles.

- Loi normale

$$[\hat{R}_i - 2se(\hat{R}_i); \hat{R}_i + 2se(\hat{R}_i)]$$

avec $se(\hat{R}_i) = \sqrt{MSEP(\hat{R}_i)}$

- Loi lognormale

$$[exp(\mu_i - 2\sigma_i); exp(\mu_i + 2\sigma_i)]$$

avec $\sigma_i^2 = \ln(1 + (se(\hat{R}_i))^2/\hat{R}_i^2)$ et $\mu_i = \ln(\hat{R}_i) - \sigma_i^2/2$

Vérification des hypothèses

- H1) pas vérifiée en pratique par exemple à cause de changements de management au niveau de la gestion des sinistres ou alors à cause du facteur d'inflation;
- pour H2) vérification graphique à l'aide du C-C plot;
- pour H3) vérification graphique : le graphe des résidus normalisés $\hat{\epsilon}_{ij}$, $i = 1, \dots, n - j$, ne doit faire apparaître aucune structure non aléatoire.

Evaluation de l'erreur pour les ratios S/P

Une estimation du ratio S/P pour l'année d'origine i sera donné par

$$\widehat{\left(\frac{S}{P}\right)}_i = \frac{\hat{C}_{in}}{P_i}$$

et l'erreur d'estimation :

$$\widehat{MSEP} \left(\left(\frac{\widehat{S}}{P} \right)_i \right) = \left(\frac{1}{P_i} \right)^2 \widehat{MSEP}(\hat{C}_{in})$$

avec $\widehat{MSEP}(\hat{C}_{in}) = \widehat{MSEP}(\hat{R}_i)$



Méthodes stochastiques

- a) Modèle recursif de Mack
- b) Modèles stochastiques factoriels
 - Régression lognormale
 - Modèle Poissonien
- c) Distribution prédictives et VaR
 - 1) Distribution prédictives par convolution
 - 2) Distribution prédictives approchées
 - 3) Distribution prédictives par bootstrap

b) Modèles stochastiques factoriels

Idée : les données du triangle supérieur des paiements non cumulés $(x_{ij})_{i+j \leq n}$ sont les réalisations des v.a. $(X_{ij})_{i+j \leq n}$.

Hypothèse : les v.a. X_{ij} , $i, j = 0, \dots, n$ sont indépendantes.

On cherche à modéliser $\mu_{ij} = E(X_{ij})$ au moyen de certaines variables explicatives.

On utilisera des modèles paramétriques car on fera une hypothèse sur la loi des v.a. X_{ij} (Poisson, Lognormale, etc.).

Rappels

La v.a. provision pour la i -ème année d'origine est donnée par :

$$R_i = \sum_{j=n-i+1}^n X_{ij} \quad i = 1, \dots, n$$

et la v.a. provision globale

$$R = \sum_{i=1}^n R_i$$

dont la f.d.r. sera notée F_R . De plus, on a

$$E(R) = \sum_{i=1}^n \sum_{i+j>n} E(X_{ij})$$

$$Var(R) = \sum_{i=1}^n \sum_{i+j>n} Var(X_{ij})$$

Variables explicatives ?

Les variables intervenant dans la modélisation correspondent aux trois directions du triangle de liquidation.

La variable "année d'origine" est qualitative ordinaire et prend les modalités $0, 1, \dots, n$. Elle est remplacé par n variables indicatrices dont les paramètres seront notés $\alpha_1, \alpha_2, \dots, \alpha_n$ ($\alpha_0 = 0$). En général

$$1_i^\alpha = \begin{cases} 1 & \text{si } x \text{ appartient à l'année d'origine } i \\ 0 & \text{sinon} \end{cases}$$

La variable "année d'origine" est représentative de l'effet ligne : effet volume du portefeuille.

Variables explicatives ?

La variable "délai de développement" est quantitative discrète à valeurs $0, 1, \dots$ mais on la considérera comme un facteur et on la remplacera par n indicatrices dont les paramètres seront notés $\beta_1, \beta_2, \dots, \beta_n$ ($\beta_0 = 0$). En général

$$1_j^\beta = \begin{cases} 1 & \text{si } x \text{ appartient à l'année de développement } j \\ 0 & \text{sinon} \end{cases}$$

La variable "délai de développement" est représentative de l'effet colonne : effet cadence de paiement.

Variables explicatives ?

Il faut aussi prendre en compte l'effet diagonale, effet "année calendaire", pour traiter l'effet de l'inflation.

L'année calendaire ferait entrer en jeu $2n$ paramètres μ_{i+j} , $i, j = 0, \dots, n$. Afin de simplifier le modèle, une hypothèse d'inflation constante est faite et donc $\mu_{i+j} = \mu$ constant.

Le modèle est le suivant

$$\mu_{ij} = E(X_{ij}) = k(\mu, \alpha_i, \beta_j)$$

Le plus souvent on travaille avec la forme additive

$$\mu_{ij} = \mu + \alpha_i + \beta_j \text{ ou multiplicatice } \mu_{ij} = \mu \alpha_i \beta_j$$

Estimation

Les paramètres du modèle peuvent être estimés par maximum de vraisemblance (MV). Soit $\theta = (\mu, (\alpha_i)_{i=1,\dots,n}, (\beta_j)_{j=1,\dots,n})$ le vecteur des paramètres. Alors

$$\hat{\theta}_{MV} = (\hat{\mu}, (\hat{\alpha}_i), (\hat{\beta}_j))$$

Par invariance fonctionnelle de l'EMV, on déduit, par plug-in, les EMV des μ_{ij}

$$\hat{\mu}_{ij} = k(\hat{\mu}, \hat{\alpha}_i, \hat{\beta}_j)$$

et l'EMV de $E(R_i)$ et $E(R)$

$$\widehat{E(R_i)} = \sum_{j=n-i+1}^n \hat{\mu}_{ij}$$

$$\widehat{E(R)} = \sum_i \sum_j \hat{\mu}_{ij}$$

Regression lognormale

On suppose que $X_{ij} \sim LN(m_{ij}, \sigma^2)$. Dans la modélisation où les variables explicatives sont année d'origine et délai de développement, on prend

$$m_{ij} = \mu + \alpha_i + \beta_j$$

avec $\alpha_0 = \beta_0 = 0$. Si on considère

$$Y_{ij} = \ln X_{ij}$$

alors, par définition, $Y_{ij} \sim N(m_{ij}, \sigma^2)$ avec

$$E(Y_{ij}) = m_{ij} = \mu + \alpha_i + \beta_j$$

Il est équivalent d'écrire

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

avec $\epsilon_{ij} \sim N(0, \sigma^2)$.

Regression lognormale

Il en résulte

$$\begin{aligned}\mu_{ij} = E(X_{ij}) &= \exp \{m_{ij} + \sigma^2/2\} \\ &= \exp \{\mu + \alpha_i + \beta_j + \sigma^2/2\}\end{aligned}$$

$$\begin{aligned}Var(X_{ij}) &= \exp \{2m_{ij} + \sigma^2\} (\exp \{\sigma^2\} - 1) \\ &= \exp(2(\mu + \alpha_i + \beta_j + \frac{\sigma^2}{2})) (e^{\sigma^2} - 1)\end{aligned}$$

Regression lognormale

Au final, il s'agit d'estimer, à partir des éléments du triangle de liquidation, un modèle de régression linéaire classique qui s'écrit sous forme matricielle :

$$Y = M\theta + \epsilon$$

avec Y le vecteur colonne, d'ordre $t = \frac{(n+1)(n+2)}{2}$, des éléments du triangle pris ligne à ligne et ϵ le vecteur associé des erreurs. θ est le vecteur des paramètres du modèle (d'ordre $p = 2n + 1$) et M est la matrice de régression dont la première colonne est le vecteur unitaire, puis les autres colonnes sont les valeurs prises par chacune des variables explicatives du modèle.

$$Y = M\theta + \varepsilon$$

$$Y = (Y_{00}, Y_{01}, \dots, Y_{0m}, Y_{10}, Y_{11}, \dots, Y_{1,m-1}, \dots, Y_{mo})'$$

$$\varepsilon = (\varepsilon_{00}, \varepsilon_{01}, \dots, \varepsilon_{mo})'$$

$$\theta = (\mu, \alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_m)$$

$$M = \begin{pmatrix} 1 & 0 & & \\ 1 & 0 & & \\ 1 & 0 & & \\ | & & & \\ 0 & & & \\ \vdots & & & \\ 1 & 0 & & \end{pmatrix}$$

M fois

m fois

1_1^α 1_2^α 1_m^β

$$\mu_{ij} = \mu + \alpha_1 1_1^\alpha + \dots + \alpha_m 1_m^\alpha + \beta_1 1_1^\beta + \dots + \beta_m 1_m^\beta$$

Regression lognormale

On procède donc comme pour le modèle linéaire classique, i.e. on estime θ par

$$\hat{\theta} = (M' M)^{-1} M' Y$$

et on obtient $\hat{\epsilon} = y_i - \hat{y}_i$ et aussi

$$\hat{\sigma}^2 = \frac{1}{t-p} \sum_{i=1}^n \hat{\epsilon}_i^2$$

ce qui permet d'obtenir

$$\hat{\mu}_{ij} = \exp\{\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\sigma}^2/2\}$$

et au final

$$\hat{R} = \sum_{i+j>n} \hat{\mu}_{ij}$$

Le modèle linéaire généralisé

On a trois composantes :

- 1) composante aléatoire du modèle : les v.a. réponses X_{ij} dont la densité appartient à la famille exponentielle :

$$f(x_{ij}; \theta_{ij}, \phi) = \exp \left\{ \frac{\theta_{ij}x_{ij} - b(\theta_{ij})}{\phi/\omega_{ij}} + c(x_{ij}, \phi) \right\}$$

- 2) composante déterministe : les variables explicatives du modèle

$$\eta_{ij} = \mu + \alpha_i + \beta_j \quad i, j = 0, \dots, n$$

avec $\alpha_0 = \beta_0 = 0$

- 3) la fonction lien g (fonction réelle, strictement monotone et derivable) tq

$$g(\mu_{ij}) = \eta_{ij}$$



Le modèle linéaire généralisé

De plus, on a que

$$\mu_{ij} = E(X_{ij}) = b'(\theta_{ij})$$

$$Var(X_{ij}) = \phi b''(\theta_{ij}) = \phi V(\mu_{ij})$$

où $V(\mu_{ij})$ est appelé fonction variance.

Modèle Poissonien (Renshaw et Verrall)

On suppose que

$$\forall i, j, \quad X_{ij} \sim P(\mu_{ij})$$

et on choisit comme fonction lien la fonction lien ln. Le modèle est donc :

$$\ln \mu_{ij} = \mu + \alpha_i + \beta_j$$

avec $\alpha_0 = \beta_0 = 0$.

Remarque : les provisions estimées dans ce modèle coincident exactement avec celle de la méthode CL standard.

Modèle de Poisson surdispersé

Il s'agit d'un modèle qui présente un paramètre supplémentaire $\phi > 0$. Il est donc plus flexible que le modèle de Poisson.

Définition $X \sim P_{surd}(\lambda, \phi)$ ssi $\frac{X}{\phi} \sim P(\lambda/\phi)$

Il en résulte

$$E(X) = \lambda$$

$$Var(X) = \phi\lambda = \phi E(X)$$



Méthodes stochastiques

- a) Modèle recursif de Mack
- b) Modèles stochastiques factoriels
 - Régression lognormale
 - Modèle Poissonien
- c) Distribution prédictives et VaR
 - 1) Distribution prédictives par convolution
 - 2) Distribution prédictives approchées
 - 3) Distribution prédictives par bootstrap



c) Distribution prédictives et VaR

Pour estimer la loi de R , dite loi prédictive, on dispose de trois possibilités :

- 1) par convolution
- 2) en se basant sur les moments de R
- 3) par des techniques de bootstrap

Rappel

Pour $1 - \eta$ fixé, le quantile d'ordre $1 - \eta$ de R est défini par

$$q_{1-\eta}(R) = F_R^{-1}(1 - \eta) = \inf\{x \in \mathbb{R} \mid F_R(x) \geq 1 - \eta\}$$

Ce quantile est aussi la Value at Risk d'ordre η de R

$$VaR_\eta(R) = q_{1-\eta}(R)$$

La $VaR_\eta(R)$ peut être interprétée comme la provision suffisante dans $100(1-\eta)\%$ des cas.

1) Distribution prédictives par convolution

Exemple: $X_{ij} \sim P(\mu_{ij})$. Alors

$$R = \sum_{i=1}^n \sum_{j=n-i+1}^n X_{ij} \sim P(\mu_R)$$

avec $\mu_R = \sum_i \sum_j \mu_{ij}$. Si $\mu_R \geq 50$,

$$R \sim_{approx} N(\mu_R, \mu_R)$$

et donc

$$P(R \leq r) \approx \Phi\left(\frac{r - \mu_R}{\sqrt{\mu_R}}\right)$$

et le quantile d'ordre $1 - \eta$ de R

$$q_{1-\eta}(R) \approx q_{1-\eta}^{(P)}(R) = \mu_R + \sqrt{\mu_R} q_{1-\eta}$$

avec $q_{1-\eta}$ le quantile d'ordre $1 - \eta$ d'une $N(0, 1)$.

On obtient un estimateur "plug-in" pour le quantile de R , $\hat{q}_{1-\eta}^{(P)}$. Simplement en remplaçant dans l'expression du quantile les quantités inconnues par leur estimateur du MV

$$\hat{q}_{1-\eta}^{(P)}(R) = \hat{\mu}_R + q_{1-\eta} \sqrt{\hat{\mu}_R}$$

estimateur du MV \rightarrow

1) Distribution prédictives par convolution

Il suffit d'estimer μ_{ij} et μ_R par MV pour obtenir un estimateur MV de $q_{1-\eta}^{(P)}(R)$.

On pourrait aussi passer par la f.g.m. $M_R(s) = E(e^{sR})$

$$M_R(s) = \prod_{i=1}^n M_{R_i}(s) = \prod_{i=1}^n \prod_{i+j>n} M_{X_{ij}}(s)$$

ou la fonction génératrice des cumulants (f.g.c.)

$$C_R(s) = \ln M_R(s)$$

Il reste à inverser la f.g.m. par la méthode FFT pour obtenir la f.d.r. de R .

FFT: FAST FOURIER TRANSFORM

2) Distribution prédictives approchées

Exemple : Approximation Normal Power (NP)

Si le coefficient d'asymétrie $\gamma_1 > 0$,

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \xleftarrow{\text{moment d'ordre 3}}$$

ϕ : fdr d'une $\mathcal{N}(0,1)$

$$F_R(x) \approx F_R^{(NP)}(x) = \Phi \left[-\frac{3}{\gamma_1} + \sqrt{\frac{9}{\gamma_1^2} + 1 + \frac{6}{\gamma_1} \left(\frac{x - \mu}{\sigma} \right)} \right]$$

pour $x > \mu - \sigma \left(\frac{\gamma_1}{\sigma} + \frac{3}{2\gamma_1} \right)$. Et aussi

$$q_{1-\eta}^{(NP)} = \mu + \sigma \left[\frac{\gamma_1}{\sigma} q_{1-\eta}^2 + q_{1-\eta} - 1 \right]$$

Remarque : comme toute approximation normale, cette approximation sous-estime la queue de distribution et donc les quantiles !

Remarque : L'approximation NP est assez précise tant que $0 < \gamma_1 \leq 2$; la qualité de l'approximation décroît à mesure que γ_1 augmente.

3) Distribution prédictives par bootstrap

Généralités On considère une v.a. X dont la f.d.r., notée F , est inconnue. On dispose d'un échantillon de X , (X_1, \dots, X_n) i.i.d.. Un estimateur sans biais et convergent de $F(x)$ est donné par la f.d.r. empirique

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{]-\infty, x]}(X_i)$$

On souhaite estimer le paramètre $\pi(F)$ et on note l'estimateur $T_n = T(X_1, \dots, X_n)$.

Un échantillon bootstrap est un n-échantillon de la f.d.r. empirique F_n . On le note $X^* = (X_1^*, \dots, X_n^*)$ tq $P(X_i^* = X_j) = 1/n$ pour $1 \leq j \leq n$.

L'estimateur bootstrap de la moyenne est simplement une moyenne empirique des estimateurs de la moyenne obtenu sur chaque réplique bootstrap.

© Théo Jalabert

$$(B=1000) \quad \frac{1}{1000} \sum_{b=1}^{1000} \bar{X}_m^{*b}$$

$$(\bar{X}_1, \dots, \bar{X}_{50})$$

$$(\bar{X}_1^{*1}, \dots, \bar{X}_{50}^{*1}) \rightarrow \bar{X}_m^{*1}$$

$$\vdots$$
$$(\bar{X}_1^{*100}, \dots, \bar{X}_{50}^{*100}) \rightarrow \bar{X}_m^{*100}$$



En pratique, on obtient une réalisation de l'échantillon bootstrap

$$x^* = (x_1^*, \dots, x_n^*)$$

en effectuant un tirage avec remise de n éléments dans l'échantillon initial (x_1, \dots, x_n) .

$T(X_1^*, \dots, X_n^*)$ est appelé replication bootstrap de T_n .

Fonctionnement du bootstrap

On tire de manière indépendante B échantillons bootstrap de F_n que l'on note

$$(X_1^{*b}, \dots, X_n^{*b}), \quad b = 1, \dots, B$$

La b -ième réPLICATION bootstrap de T_n est

$$T_n^{*b} = T(X_1^{*b}, \dots, X_n^{*b}), \quad b = 1, \dots, B$$

puis l'estimateur bootstrap de $\pi(F)$ est

$$\bar{T}_n^* = \frac{1}{B} \sum_{b=1}^B T_n^{*b}$$

Une approximation Monte-Carlo de la variance bootstrap de T_n est obtenue par la variance empirique de ces B réPLICATIONS

$$V_{Boot}^{(B)} = \frac{1}{B-1} \sum_{b=1}^B (T_n^{*b} - \bar{T}_n^*)^2$$

Un estimateur bootstrap de $MSE(T_n)$ serait

$$MSE_{Boot}^{(B)} = \frac{1}{B} \sum_{b=1}^B [T(X_1^{*b}, \dots, X_n^{*b}) - \pi(F_n)]^2$$

et pour le biais $B(T_n) = E(T_n) - \pi(F)$, on obtient l'estimateur bootstrap

$$B_{Boot}^{(B)} = \frac{1}{B} \sum_{b=1}^B T(X_1^{*b}, \dots, X_n^{*b}) - \pi(F_n)$$

Distribution d'échantillonage de T_n

Soit H_n la f.d.r. de T_n

$$H_n(x) = P(T_n \leq x) = P(T(X_1, \dots, X_n) \leq x)$$

L'estimateur bootstrap de $H_n(x)$ est donné par

$$H_{Boot}(x) = P(T_n^* \leq x \mid X_1, \dots, X_n)$$

où $T_n^* = T(X_1^*, \dots, X_n^*)$ pour un échantillon bootstrap (X_1^*, \dots, X_n^*) .

Mais malheureusement il n'est pas toujours ais  de calculer analytiquement $H_{Boot}(x)$!

Distribution d'échantillonage de T_n

Dans ce cas, on lui substitue l'approximation Monte-Carlo

$$H_{Boot}^{(B)}(x) = \frac{1}{B} \sum_{b=1}^B \mathbf{1}_{\{T_n^{*b} \leq x\}}$$

avec $T_n^{*b} = T(X_1^{*b}, \dots, X_n^{*b})$ la b -ième réPLICATION bootstrap et $(X_1^{*b}, \dots, X_n^{*b})_{b=1, \dots, B}$ un ensemble de B échantillons bootstrap (indépendants).

Un intervalle de confiance pour $\pi(F)$ de niveau asymptotique $1 - \eta$ peut être obtenu par

$$[H_{Boot}^{-1}(\eta/2), H_{Boot}^{-1}(1 - \eta/2)]$$

Le bootstrap en provisionnement !

Problème : on ne peut pas appliquer le bootstrap directement sur les éléments du triangle supérieur, x_{ij} , $i + j \leq n$ car les v.a. X_{ij} sont indépendantes mais pas identiquement distribuées.

Idée : on ajuste un modèle pertinent aux X_{ij} , on calcule les résidus du modèle

$$r_{ij} = h(x_{ij}, \hat{\mu}_{ij})$$

et on procède ensuite au rééchantillonage des résidus $(r_{ij})_{i+j \leq n}$.

Mais allons regarder plus en détail les étapes de la procédure bootstrap !

Étapes de la procédure bootstrap

- 1) on estime les paramètres du modèle de régression, ce qui permet d'obtenir les valeurs prévues par le modèle $\hat{\mu}_{ij} = g^{-1}(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)$ et le best estimate

$$\widehat{E(R)} = \hat{R} = \sum_{i+j>n} \hat{\mu}_{ij}$$

- 2) On calcule les résidus de Pearson (r_{ij}^P) , $i + j \leq n$, par

$$r_{ij}^P = \frac{x_{ij} - \hat{\mu}_{ij}}{\sqrt{V(\hat{\mu}_{ij})}}$$

et on estime (si nécessaire) le paramètre de dispersion ϕ à l'aide des résidus de Pearson

$$\hat{\phi} = \frac{1}{t-p} \sum_{i+j \leq n} (r_{ij}^P)^2$$

On peut maintenant commencer le bootstrap :

3) on répète B fois ($b = 1, \dots, B$) les opérations suivantes:

- on obtient un échantillon bootstrap de résidus $(r_{ij}^{*b})_{i+j \leq n}$ par rééchantillonage des résidus initiaux du modèle
- on détermine le triangle d'increments bootstrappés associés $(x_{ij}^{*b})_{i+j \leq n}$ par

$$x_{ij}^{*b} = \hat{\mu}_{ij} + r_{ij}^{*b} \sqrt{V(\hat{\mu}_{ij})}$$

- à partir de ces nouvelles données de triangle supérieur, on obtient à nouveau les valeurs prévues $(\hat{\mu}_{ij}^{*b})_{i+j > n}$ et on calcule le best estimate $\hat{R}^{*b} = \sum_{i+j > n} \hat{\mu}_{ij}^{*b}$
- on stocke \hat{R}^{*b} et on répète le bootstrap

4) on peut maintenant utiliser le B - échantillon bootstrap $(\hat{R}^{*1}, \dots, \hat{R}^{*B})$ pour calculer par exemple les quantités suivantes :

- $E_{boot}(\hat{R}) = \frac{1}{B} \sum_{b=1}^B \hat{R}^{*b}$
- le risque d'estimation

$$MSE_{boot}(\hat{R}) = \frac{1}{B-1} \sum_{b=1}^B [\hat{R}^{*b} - E_{boot}(\hat{R})]^2$$

- le risque de processus

$$Var_{boot}(R) = \hat{\phi} \sum_{i+j>n} V(\hat{\mu}_{ij})$$

- l'erreur de prédiction

$$sep_{boot}(\hat{R}) = \sqrt{MSEPC_{boot}(\hat{R})} = \sqrt{MSE_{boot}(\hat{R}) + Var_{boot}(R)}$$

Limites de l'approche bootstrap

- le bootstrap ne s'applique que sur des résidus i.i.d. et cette hypothèse doit être vérifiée;
- le bootstrap n'est pas robuste pour l'estimation de quantiles élevés dans la mesure où souvent les données que l'on possède ne reflètent pas les observations concernant ces quantiles.