

Cours 4 – Bootstrap : les aspects pratiques

Bruno PORTIER

INSA – GM5 – Cours de Statistique
Régression Non Linéaire avec Applications sous R

2006-2007

Quelques références bibliographiques

- Efron, B. and Tibshirani, R. (1986). *The bootstrap method for standard errors, confidence intervals, and other measures of statistical accuracy*. Statistical Science, Vol 1., No. 1, pp 1-35.
- Efron, B. (1992) *Jackknife-after-bootstrap standard errors and influence functions*. Journal of . Roy. Stat. Soc. B, vol 54, pages 83-127
- Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, New York, London.
- Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer.
- Davison, A. C., Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge : Cambridge University Press.
- Good, P. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Berlin : Springer

Ce cours est très largement inspiré de l'article de Rudy Palm
(Faculté de Gembloux)

<http://www.bib.fsagx.ac.be/library/base/summary/v6n3/143.pdf>

et des supports de cours de

- Irène Buvat (U494 INSERM)

<http://guillemet.org/irene/coursem/bootstrap.pdf>

- Christian P. Robert (Université Paris-Dauphine)

<http://www.ceremade.dauphine.fr/~xian/Noise/Chap3.pdf>

1. Introduction

- Les méthodes classiques de la statistique inférentielle ne permettent pas toujours d'obtenir des réponses correctes aux problèmes qu'on peut rencontrer dans la pratique.
- En effet, ces méthodes
 - dépendent souvent de l'hypothèse de normalité
 - ou bien s'appuient sur un résultat asymptotique et ne peuvent donc être utilisées que lorsque la taille de l'échantillon est "suffisamment grande"ce qui n'est pas toujours le cas en pratique.

Que peut-on faire alors, en pratique, lorsque les conditions d'application ne sont pas remplies ?

Différentes attitudes sont possibles.

- Appliquer la méthode sans se poser la moindre question (fortement déconseillé) ;
- Appliquer la méthode, malgré le non-respect des conditions, parce qu'on sait qu'elle est robuste, ie. la méthode garantit que les résultats de l'inférence restent approximativement valables. C'est le cas de l'ANOVA.
- Recourir à des transformations de variables permet, dans certains cas, de se rapprocher des conditions d'application. Ainsi une transformation logarithmique, par exemple, peut rendre normales des distributions qui, au départ, ne le sont pas.
- Utiliser le **bootstrap**

1.1 Qu'est-ce que le bootstrap ?

Le **Bootstrap**, c'est une

- Technique récente (1979) qui a fait son apparition avec le développement des moyens de calcul informatiques.
- Technique basée sur le principe du rééchantillonnage, .
- Technique permettant de faire de l'inférence statistique à partir d'un nombre limité d'observations en utilisant une succession de rééchantillonnage.
- Technique qui facilite l'inférence statistique dans les situations complexes où les méthodes analytiques ne suffisent pas

1.2 Que va nous permettre de faire le bootstrap ?

Le **bootstrap** va nous permettre, lorsque les méthodes classiques de la statistique ne s'appliquent pas, parce que

- les hypothèses de normalité ne sont pas vérifiées
- la taille de l'échantillon est trop petite pour pouvoir utiliser l'approximation fournie par le TLC

de résoudre des problèmes usuels, comme

- réduire le biais d'un estimateur
- étudier la variabilité empirique d'estimateurs
- construire des intervalles de confiance et/ou de prévision empiriques
- tester des hypothèses

Cependant, les “solutions bootstrap”

- ne sont pas destinées à remplacer les méthodes d'inférence statistique classiques lorsque celles-ci sont applicables

mais

- sont plutôt destinées à fournir des réponses à des questions pour lesquelles les méthodes classiques sont inapplicables ou non disponibles.

1.3 En quoi consiste le rééchantillonnage ?

On distingue deux méthodes de rééchantillonnage bootstrap :

- le rééchantillonnage basé sur les individus

et lorsque celui-ci ne peut pas être justifié

- le rééchantillonnage basé sur les résidus

1.3.1 Bootstrap des individus

On considère un échantillon de n observations :

$$X_1, X_2, \dots, X_i, \dots, X_n$$

prélevé de manière aléatoire dans une population.

Ces observations peuvent concerner une seule variable, ou, au contraire, être relatives à plusieurs variables.

Le principe de la méthode du **bootstrap** est d'effectuer une succession de **tirages aléatoires avec remise** de n valeurs parmi les n valeurs de l'échantillon initial.

Exemple.

- Un échantillon observé de $n = 10$ valeurs

$$x = (50, 53, 58, 80, 75, 69, 77, 44, 63, 73)$$

- Un premier échantillon bootstrap :

$$x^{*1} = (77, 73, 69, 44, 77, 50, 44, 53, 58, 69)$$

- Un deuxième échantillon bootstrap :

$$x^{*2} = (63, 75, 63, 80, 69, 44, 44, 73, 63, 77)$$

1.3.2 Bootstrap des résidus

Lorsque le rééchantillonnage des individus ne peut se justifier,

- comme par exemple dans le cas de la régression linéaire où les valeurs des variables explicatives sont fixées a priori par l'utilisateur,

on remplace le *bootstrap des individus* par le *bootstrap des résidus*.

Soient

- Y le vecteur de la variable à expliquer
- et X la matrice des variables explicatives.

Supposons qu'on envisage une liaison du type

$$Y = f(X, \theta) + \varepsilon$$

où

- θ est le vecteur des paramètres
- ε est un vecteur centré, de loi inconnue

Soit $\hat{\theta}$ un estimateur de θ . On peut alors

- prédire le vecteur Y par $\hat{Y} = f(X, \hat{\theta})$.
- définir le vecteur des résidus $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)'$ par

$$\hat{\varepsilon} = Y - \hat{Y} = Y - f(X, \hat{\theta})$$

Construction d'un échantillon bootstrap.

- 1 Si le vecteur $\hat{\varepsilon}$ n'est pas centré, on le centre.
- 2 On tire au hasard avec remise dans le vecteur $\hat{\varepsilon}$, n résidus ($\hat{\varepsilon}_i$) que l'on note

$$(\hat{\varepsilon}_1^{(*)}, \dots, \hat{\varepsilon}_n^{(*)}) = \hat{\varepsilon}^{(*)}$$

- 3 On construit une réplique bootstrap de Y

$$Y^{(*)} = f(X, \hat{\theta}) + \hat{\varepsilon}^{(*)}$$

L'échantillon bootstrap est alors donné par $(X, Y^{(*)})$.

2. Justification de la méthode de rééchantillonnage

Soient X_1, X_2, \dots, X_n des variables aléatoires réelles indépendantes et identiquement distribuées de loi F inconnue.

On dispose d'une réalisation (x_1, x_2, \dots, x_n) de (X_1, X_2, \dots, X_n) .

Comment construire une autre réalisation de (X_1, X_2, \dots, X_n) ?

- Aucun problème lorsqu'on connaît F : il suffit de simuler n valeurs d'une variable aléatoire de loi F .

Par exemple, lorsque F^{-1} est explicite, on simule une valeur x de X

- en simulant une valeur u de $U \sim \mathcal{U}_{[0,1]}$
- et en calculant $x = F^{-1}(u)$.

- Lorsque F est inconnue, on estime F par la fonction de répartition empirique F_n définie pour tout $x \in \mathbb{R}$ par

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{\{X_k \leq x\}}.$$

et on simule n valeurs selon la loi F_n .

On sait, d'après le théorème de Glivenko-Cantelli, que pour tout $x \in \mathbb{R}$,

$$F_n(x) \xrightarrow[n \rightarrow \infty]{p.s.} F(x)$$

et que cette convergence est uniforme sur \mathbb{R} , ie.

$$\|F_n - F\|_{\infty} := \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n \rightarrow \infty]{p.s.} 0$$

Soient $X_1^*, X_2^*, \dots, X_n^*$ des variables aléatoires réelles indépendantes et identiquement distribuées de loi F_n .

Le vecteur $X^* = (X_1^*, X_2^*, \dots, X_n^*)$, obtenu par rééchantillonnage, est appelé échantillon bootstrap.

De plus, pour tout $1 \leq i, j \leq n$, on a

$$\mathbb{P} [X_j^* = X_i \mid X_1, \dots, X_n] = \frac{1}{n}$$

Par conséquent, à partir de la réalisation x_1, x_2, \dots, x_n , on peut construire une estimation de F et se servir de cette estimation pour construire une autre réalisation $x_1^*, x_2^*, \dots, x_n^*$,

Concrètement, puisque la fonction de répartition empirique estimée attribue à chaque x_i une probabilité $1/n$ d'être tiré, pour simuler n valeurs selon la loi F_n , il suffit de tirer avec remise n valeurs parmi les n valeurs de l'échantillon x_1, x_2, \dots, x_n .

Cette méthode constitue ce qu'on appelle le **bootstrap non-paramétrique**.

3. Estimation de l'erreur standard et du biais

On considère une variable aléatoire X de loi F inconnue. Soient X_1, X_2, \dots, X_n des variables aléatoires iid de même loi que X . Soit F_n la fonction de répartition empirique associée :

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{\{X_k \leq x\}}.$$

On s'intéresse à un paramètre $\theta \in \mathbb{R}$ de la loi F .

Le paramètre θ peut s'écrire comme une fonctionnelle de F , soit $\theta = t(F)$.

Un estimateur naturel de $\theta = t(F)$ est alors donné par

$$\hat{\theta} = t(F_n) = T(X_1, \dots, X_n)$$

Par exemple,

❶ si $\theta = E[X] = \int_{\mathbb{R}} x dF(x)$ alors $\hat{\theta} = \int_{\mathbb{R}} x dF_n(x) = \frac{1}{n} \sum_{i=1}^n X_i$.

❷ si $\theta = E[h(X)]$ où h est une fonction de \mathbb{R} dans \mathbb{R} , alors

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n h(X_i)$$

❸ si θ est la médiane de X , alors

$$\hat{\theta} = \begin{cases} \frac{X_{([n/2])} + X_{([n/2]+1)}}{2} & \text{si } n \text{ est pair} \\ X_{((n+1)/2)} & \text{si } n \text{ est impair} \end{cases}$$

où $X_{(1)}, \dots, X_{(n)}$ désigne la statistique d'ordre associée à X_1, \dots, X_n .

❹ si θ est le quantile de niveau $(1 - \alpha)$ de la loi de X , alors $\hat{\theta} = X_{([(1-\alpha)n]+1)}$.

On s'intéresse alors aux problèmes d'estimation suivants, qui sont à l'origine du Bootstrap :

- estimer le biais de $\hat{\theta}$
- estimer l'écart-type encore appelé erreur standard de $\hat{\theta}$
- estimer l'erreur quadratique moyenne de $\hat{\theta}$

à partir d'une seule réalisation de (X_1, X_2, \dots, X_n) .

3.1 Estimation du biais par Bootstrap

Objectif. On veut estimer, sur la base d'une observation de X_1, \dots, X_n , le biais de $\hat{\theta}$.

Ce biais est défini par :

$$b(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

Ce biais est inconnu puisqu'il dépend de la loi F inconnue.

Comment l'estimer ?

Supposons F et θ connus (*hypothèse d'école*)

- Soit $b(\hat{\theta})$ se calcule de manière analytique et c'est fini.
- Soit $b(\hat{\theta})$ ne se calcule pas de manière analytique et on l'approxime alors par Monte-Carlo.

Cette méthode repose sur la loi forte des grands nombres.

On sait que si $(\hat{\theta}_1, \dots, \hat{\theta}_B)$ est un B -échantillon de la loi de $\hat{\theta}$ et si $\mathbb{E}[|\hat{\theta}|] < \infty$, alors

$$\frac{1}{B} \sum_{\ell=1}^B \hat{\theta}_{\ell} \xrightarrow[n \rightarrow \infty]{p.s.} \mathbb{E}[\hat{\theta}]$$

Méthode de Monte-Carlo (*pour le calcul du biais*)

Pour $\ell = 1, \dots, B$ (B grand)

- ❶ on simule un n -échantillon $(X_1^\ell, \dots, X_n^\ell)$ de loi F
- ❷ on calcule $\hat{\theta}_\ell = T(X_1^\ell, \dots, X_n^\ell)$

On estime finalement $b(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$ par :

$$\widehat{b(\hat{\theta})} = \frac{1}{B} \sum_{j=1}^B \hat{\theta}_j - \theta$$

La loi F et θ sont inconnus.

On estime le biais $b(\hat{\theta})$ par la méthode du bootstrap qui consiste, dans l'algorithme de Monte Carlo précédent,

- à remplacer θ par son estimation $\hat{\theta}$
- et à simuler des échantillons selon la loi de F_n au lieu de F .

Méthode du Bootstrap (*pour le calcul du biais*)

Pour $\ell = 1, \dots, B$ (B grand)

- 1 on construit un échantillon bootstrap $(x_1^{*\ell}, x_1^{*\ell}, \dots, x_n^{*\ell})$ en effectuant un tirage aléatoire avec remise de n valeurs dans l'échantillon initial x_1, x_2, \dots, x_n
- 2 on calcule $\hat{\theta}_\ell^* = T(x_1^{*\ell}, x_1^{*\ell}, \dots, x_n^{*\ell})$

On estime finalement $b(\hat{\theta})$ par : $\widehat{b(\hat{\theta})} = \frac{1}{B} \sum_{j=1}^B \hat{\theta}_j^* - \hat{\theta}.$

Illustration

On s'intéresse à la concentration maximale journalière en SO_2 dans la région Rouennaise. On dispose des mesures de l'année 1995.

**Maximum de la concentration journalière de SO_2
(Région Rouennaise, Année 1995)**

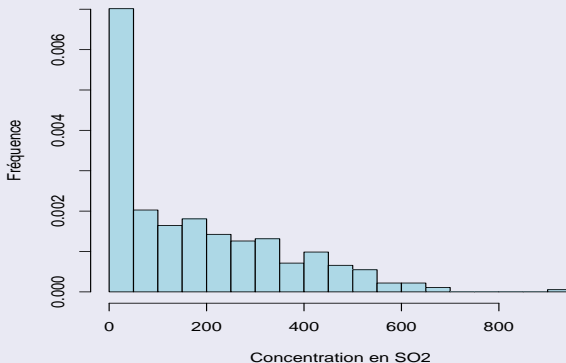


Illustration (*suite*)

On s'intéresse à la moyenne et à la médiane de la série.

On trouve $m = 181.67$ et $me = 133$.

Ces valeurs sont-elles biaisées ou pas ?

On estime le biais par bootstrap. On trouve

B	50	100	200	500	1000	2000
$b(m)$	2.227	0.612	-0.653	-0.096	-0.182	0.098
$b(me)$	0.6	-1.81	-3.55	-2.772	-2.757	-2.334

3.2 Estimation de l'erreur standard de $\hat{\theta}$ par Bootstrap

Objectif. On veut estimer, sur la base de l'observation de X_1, \dots, X_n , l'écart-type ou erreur standard de $\hat{\theta}$.

Elle est définie par :

$$SE(\hat{\theta}) = \sqrt{\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]}$$

Supposons F connue.

- Soit $SE(\hat{\theta})$ se calcule de manière analytique et c'est fini.
- Soit $SE(\hat{\theta})$ ne se calcule pas de manière analytique et on l'approxime alors par Monte-Carlo.

Méthode de Monte-Carlo (*pour le calcul de l'erreur standard*)

Pour $\ell = 1, \dots, B$ (B grand)

- ① on simule un n -échantillon $(X_1^\ell, \dots, X_n^\ell)$ de loi F
- ② on calcule $\hat{\theta}_\ell = T(X_1^\ell, \dots, X_n^\ell)$

On estime finalement $SE(\hat{\theta})$ par :

$$\widehat{SE(\hat{\theta})} = \sqrt{\frac{1}{B-1} \sum_{\ell=1}^B \left(\hat{\theta}_\ell - \frac{1}{B} \sum_{j=1}^B \hat{\theta}_j \right)^2}$$

La loi F est inconnue et on ne dispose que d'une seule réalisation x_1, \dots, x_n de X_1, \dots, X_n . On utilise donc pour estimer l'erreur standard $SE(\hat{\theta})$ la méthode du bootstrap (des individus).

Méthode du Bootstrap (*pour le calcul de l'erreur standard*)

Pour $\ell = 1, \dots, B$ (B grand)

- 1 on construit un échantillon bootstrap $(x_1^{*\ell}, x_1^{*\ell}, \dots, x_n^{*\ell})$ en effectuant un tirage aléatoire avec remise de n valeurs dans l'échantillon initial x_1, x_2, \dots, x_n
- 2 on calcule $\hat{\theta}_\ell^* = T(x_1^{*\ell}, x_1^{*\ell}, \dots, x_n^{*\ell})$

On estime finalement $SE(\hat{\theta})$ par :

$$\widehat{SE(\hat{\theta})} = \sqrt{\frac{1}{B-1} \sum_{\ell=1}^B \left(\hat{\theta}_\ell^* - \frac{1}{B} \sum_{j=1}^B \hat{\theta}_j^* \right)^2}$$

Exemple (*Durée de Survie de souris* (Efron))

Lors d'une petite expérimentation sur des souris atteintes d'une maladie mortelle, on a tiré au sort parmi 16 souris, 7 qui reçoivent un nouveau traitement alors que les 9 autres sont des contrôles qui reçoivent un placebo. Leurs durées de survie sont mesurées en jours et donnent les résultats suivants :

Survie (en jours)

Groupe 1 (Placebo)

$n_1 = 9$ mesures

52, 10, 40, 104, 50,
27, 146, 31, 46

moyenne $m_1 = 56.22$

médiane $\mu_1 = 46$

Groupe 2 (Traitement)

$n_2 = 7$ mesures

94, 38, 23, 197,
99, 16, 141

moyenne $m_2 = 86.86$

médiane $\mu_2 = 94$

Exemple (*suite et fin*)

On a l'impression que le traitement assure une meilleure survie que le placebo.

Mais les échantillons sont petits et la précision des deux estimations des moyennes réelles est certainement très mauvaise.

Comment mesurer cette précision ? **Bootstrap**

Erreur standard sur m_1 :

↪ classique : $se_1 = 14.14$

↪ bootstrap : $se_1^* = 13.32$

Erreur standard sur μ_1 :

↪ classique : ?

↪ bootstrap : $se_1^* = 11.54$

Erreur standard sur m_2 :

↪ classique : $se_2 = 25.24$

↪ bootstrap : $se_2^* = 23.81$

Erreur standard sur μ_2 :

↪ classique : ?

↪ bootstrap : $se_2^* = 36.35$

3.3 Estimations par *jackknife*

Une autre forme de rééchantillonnage permettant d'estimer l'erreur-standard et le biais d'un estimateur est la technique du **jackknife**.

vspac1ex

L'idée de cette méthode de rééchantillonnage est de construire, à partir de l'échantillon initial et en enlevant à chaque fois une observation différente, n échantillons de taille $(n - 1)$.

On peut alors bâtir n estimations du paramètre θ , que l'on notera

$$\hat{\theta}_{(-1)}, \dots, \hat{\theta}_{(-i)}, \dots, \hat{\theta}_{(-n)}$$

l'estimation $\hat{\theta}_{(-i)}$ étant construite à partir de l'échantillon initial duquel on aura retiré la donnée x_i .

On note $\hat{\theta}_J = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(-i)}$.

On estime alors

- le biais par :

$$(n-1) (\hat{\theta}_J - \hat{\theta})$$

- l'erreur standard par :

$$\sqrt{\frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{(-i)} - \hat{\theta}_J)^2}$$

4. Construction d'intervalles de confiance

Soient X_1, X_2, \dots, X_n des variables aléatoires iid de loi inconnue F ,
et soient x_1, x_2, \dots, x_n une réalisation de X_1, X_2, \dots, X_n .

Soit F_n la fonction de répartition empirique associée :

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{\{X_k \leq x\}}.$$

On s'intéresse à un paramètre $\theta = t(F) \in \mathbb{R}$ de la loi F .

Un estimateur naturel de $\theta = t(F)$ est alors donné par

$$\hat{\theta} = t(F_n) = T(X_1, \dots, X_n)$$

Un intervalle de confiance pour le paramètre θ au niveau $(1 - \alpha)$, est un intervalle de la forme

$$[b_1(x_1, \dots, x_n); b_2(x_1, \dots, x_n)]$$

avec

$$\mathbb{P}\left[\theta \in [b_1(X_1, \dots, X_n); b_2(X_1, \dots, X_n)]\right] \geq 1 - \alpha$$

et où les variables aléatoires $b_1(X_1, \dots, X_n)$ et $b_2(X_1, \dots, X_n)$ ne dépendent pas de F .

On note alors

$$\text{IC}_{1-\alpha}(\theta) = [b_1(x_1, \dots, x_n); b_2(x_1, \dots, x_n)]$$

4.1. Méthode de l'erreur-standard

Une première solution consiste à définir l'intervalle de confiance par la méthode de l'erreur-standard (*standard bootstrap confidence interval*) : au niveau $(1 - \alpha)$, on a

$$IC_{1-\alpha}(\theta) = \left[\hat{\theta} \pm u_{1-\alpha/2} \hat{\sigma}_{\hat{\theta}*} \right]$$

où

- $u_{1-\alpha/2}$ est le quantile $(1 - \alpha/2)$ de la loi normale centrée réduite, ie. $u_{1-\alpha/2}$ est tel que

$$\mathbb{P} \left[|\mathcal{N}(0, 1)| \leq u_{1-\alpha/2} \right] = 1 - \alpha \iff \mathbb{P} \left[\mathcal{N}(0, 1) \leq u_{1-\alpha/2} \right] = 1 - \frac{\alpha}{2}$$

- $\hat{\sigma}_{\hat{\theta}*}$ est une estimation de l'erreur standard de $\hat{\theta}$.

Pour que cette approche soit satisfaisante, il faut que

- la distribution d'échantillonnage du paramètre étudié soit approximativement normale : cela peut être vérifiée à partir de la distribution des $(\hat{\theta}_k^*)$.
- l'estimateur soit non biaisé : on peut éventuellement assurer le débiaisage (voir paragraphe 3.2) en prenant le risque d'augmenter la variance de l'estimateur.
- $\hat{\sigma}_{\hat{\theta}^*}$ soit une bonne estimation de l'erreur-standard de la distribution du paramètre : pour cela il suffit de prendre un nombre B d'échantillons bootstrap plus important.

Illustration

4.2 Méthode des pourcentiles simples

Dans la méthode des pourcentiles simples
(*simple percentile confidence interval*),
les limites de l'intervalle de confiance au niveau $(1 - \alpha)$
sont données par les pourcentiles $\alpha/2$ et $1 - \alpha/2$ de la distribution
des estimations bootstrap $(\hat{\theta}_k^*)$.

On les notera $\hat{\theta}_{[\alpha/2]}^*$ et $\hat{\theta}_{[1-\alpha/2]}^*$ et l'intervalle de confiance aura
pour expression :

$$\text{IC}_{1-\alpha}(\theta) = \left[\hat{\theta}_{[\alpha/2]}^*, \hat{\theta}_{[1-\alpha/2]}^* \right]$$

On peut noter que, contrairement à la méthode de l'erreur-standard,

- la distribution d'échantillonnage du paramètre étudié peut être quelconque ;
- le nombre B de rééchantillonnages doit être plus élevé, car il faut un plus grand nombre d'observations pour estimer, avec une précision suffisante, un pourcentile : B sera par exemple de l'ordre de 1000.

Par exemple, pour 1000 rééchantillonnages et pour un niveau de confiance de 95%, les pourcentiles 0,025 et 0,975 correspondent approximativement aux observations $\tilde{\theta}_{25}^*$ et $\tilde{\theta}_{975}^*$ où la suite $(\tilde{\theta}_k^*)$ est la suite ordonnée par ordre croissant des $(\hat{\theta}_k^*)$.

Illustration

On peut mentionner une procédure de calcul un peu différente, proposée par Hall (1992) et décrite par Manly (1997).

La méthode consiste à calculer la suite des écarts :

$$\hat{e}_k^* = \hat{\theta}_k^* - \hat{\theta}$$

et à déterminer les pourcentiles $\alpha/2$ et $1 - \alpha/2$, notés $\hat{e}_{[\alpha/2]}^*$ et $\hat{e}_{[1-\alpha/2]}^*$, de la distribution des (\hat{e}_k^*) .

L'intervalle de confiance au niveau $(1 - \alpha)$ est alors de la forme :

$$IC_{1-\alpha}(\theta) = \left[\hat{\theta} - \hat{e}_{[1-\alpha/2]}^*, \hat{\theta} - \hat{e}_{[\alpha/2]}^* \right]$$

Illustration

4.3 Méthode des pourcentiles corrigés pour le biais

On commence par calculer

- 1 la proportion p des estimations bootstrap $(\hat{\theta}_k^*)$ inférieures à $\hat{\theta}$.
- 2 le pourcentile u_p relatif à la distribution normale centrée réduite, donné par

$$\mathbb{P}[\mathcal{N}(0, 1) \leq u_p] = p$$

- 3 les valeurs α_1 et α_2 définies par

$$\alpha_1 = \Phi(2u_p + u_{\alpha/2}) \quad \text{et} \quad \alpha_2 = \Phi(2u_p + u_{1-\alpha/2})$$

où

- Φ désigne la fonction de répartition de la loi $\mathcal{N}(0, 1)$
- u_β est donné par $\mathbb{P}[\mathcal{N}(0, 1) \leq u_\beta] = \beta$.

Les limites de l'intervalle de confiance au niveau $(1 - \alpha)$, déterminées par la méthode des pourcentiles corrigés pour le biais (*bias corrected percentile confidence interval*) sont alors fournies par les pourcentiles $\hat{\theta}_{[\alpha_1]}^*$ et $\hat{\theta}_{[\alpha_2]}^*$ de la distribution des $(\hat{\theta}_k^*)$.

L'intervalle de confiance est donc défini par :

$$IC_{1-\alpha}(\theta) = \left[\hat{\theta}_{[\alpha_1]}^*, \hat{\theta}_{[\alpha_2]}^* \right]$$

Des informations concernant l'origine de cette correction sont données dans le livre d'Efron et Tibshirani (1993).

Remarque

- Si $p = 0.5$, c'est-à-dire si $\hat{\theta}$ est la médiane de la distribution des $(\hat{\theta}_k^*)$, alors il n'y a pas de correction pour le biais, puisque $u_p = 0$, et on retrouve la méthode précédente.
- Si $p < 0.5$, les limites de confiance correspondent à des pourcentiles inférieurs respectivement à $\alpha/2$ et $1 - \alpha/2$. Par exemple, si $p = 0.4$, les limites de confiance, pour un niveau de confiance de 95%, sont les pourcentiles $\hat{\theta}_{[0,0068]}^*$ et $\hat{\theta}_{[0,9269]}^*$.
- Au contraire, si $p > 0.5$, les limites correspondent à des pourcentiles supérieurs à $\alpha/2$ et $1 - \alpha/2$. Par exemple, si $p = 0.6$, les limites de confiance, pour un niveau de confiance de 95%, sont les pourcentiles $\hat{\theta}_{[0,0731]}^*$ et $\hat{\theta}_{[0,9932]}^*$.

Illustration

4.4 Méthode des pourcentiles avec correction pour le biais et accélération

La méthode *bias corrected and accelerated confidence interval* est une extension de la méthode précédente qui permet de prendre en compte un éventuel changement de l'erreur-standard de $\hat{\theta}$ lorsque θ varie. On peut trouver une justification de cette méthode dans Efron et Tibshirani (1993).

L'intervalle de confiance est ici défini par :

$$IC_{1-\alpha}(\theta) = \left[\hat{\theta}_{[\alpha_1]}^*, \hat{\theta}_{[\alpha_2]}^* \right]$$

où les valeurs α_1 et α_2 sont définies par :

$$\alpha_1 = \Phi \left(u_p + \frac{u_p + u_{\alpha/2}}{1 - a(u_p + u_{\alpha/2})} \right), \quad \alpha_2 = \Phi \left(u_p + \frac{u_p + u_{1-\alpha/2}}{1 - a(u_p + u_{1-\alpha/2})} \right)$$

où

- u_p est défini comme dans le paragraphe 4.3
- la constante a est appelée **accélération**, car elle est liée au taux de variation de l'erreur-standard de $\hat{\theta}$ lorsque le paramètre θ varie.

La constante a peut être estimée de différentes manières. Une solution consiste à utiliser la technique du *jackknife*. On obtient alors le paramètre a par la formule suivante :

$$a = \frac{\sum_{i=1}^n \left(\hat{\theta}_{(-i)} - \hat{\theta}_J \right)^3}{6 \left(\sum_{i=1}^n \left(\hat{\theta}_{(-i)} - \hat{\theta}_J \right)^2 \right)^{3/2}}$$

5. Bootstrap en régression linéaire

Objectif. Présenter la méthode du Bootstrap en régression linéaire pour l'obtention

- d'intervalle de confiance pour le vecteur des paramètres
- d'intervalle de prévision

sans faire l'hypothèse de normalité pour la loi des erreurs ε .

Dans le contexte du modèle de régression linéaire, **on rééchantillonnera les résidus.**

5.1 Le modèle

On considère le modèle de régression linéaire, écrit sous la forme vectorielle

$$Y = X\theta + \varepsilon$$

où

- $Y = (Y_1, \dots, Y_n)'$
- $\theta = (\theta_0, \theta_1, \dots, \theta_p)'$
- X est la matrice des régressuers de format $n \times (p + 1)$ et supposée de rang $(p + 1)$.
- $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ est un vecteur aléatoire centré, de matrice de variance-covariance $\sigma^2 I_n$.

Dans le cadre du modèle de régression linéaire

- on estime le paramètre θ à l'aide de l'estimateur des moindres carrés ordinaires :

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

- le vecteur des résidus est défini par

$$\hat{\varepsilon} = Y - X\hat{\theta}$$

- le vecteur $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)'$ est centré et de matrice variance-covariance $\sigma^2(I_n - X(X'X)^{-1}X')$.

5.2 Construction d'un échantillon bootstrap.

- ❶ A partir des résidus

$$\widehat{\varepsilon}_1, \widehat{\varepsilon}_2, \dots, \widehat{\varepsilon}_n$$

on tire au hasard avec remise n résidus que l'on note

$$\widehat{\varepsilon}_1^*, \widehat{\varepsilon}_2^*, \dots, \widehat{\varepsilon}_n^*$$

- ❷ A partir de ces n résidus, on construit un nouvel échantillon à l'aide de la formule

$$Y^* = X\widehat{\theta} + \widehat{\varepsilon}^*$$

appelé échantillon bootstrapé ou encore échantillon étoile.

- ❸ A partir de l'échantillon étoile (X, Y^*) , on réestime le vecteur des paramètres θ :

$$\widehat{\theta}^* = (X'X)^{-1}X'Y^*.$$

Résultat théorique. La théorie du Bootstrap indique que la distribution de

$$\sqrt{n}(\hat{\theta}^* - \hat{\theta}),$$

distribution que nous pouvons calculer directement à partir des données, approche correctement la distribution de

$$\sqrt{n}(\hat{\theta} - \theta)$$

qui elle ne peut pas être calculée.

5.3 Intervalle de Confiance et de Prévision

On commence par construire B estimations bootstrap de θ :

Répéter pour k de 1 jusqu'à B

- tirer au hasard avec remise n résidus $(\hat{\varepsilon}_i)$ notés

$$(\hat{\varepsilon}_1^{(k)}, \dots, \hat{\varepsilon}_n^{(k)})' = \hat{\varepsilon}^{(k)}$$

- à partir de ce nouveau vecteur de résidus, construire un échantillon bootstrap

$$Y^{(k)} = X\hat{\theta} + \hat{\varepsilon}^{(k)}$$

- à partir de cet échantillon bootstrapé, construire l'estimation bootstrap $\hat{\theta}^{(k)}$ de θ :

$$\hat{\theta}^{(k)} = (X'X)^{-1}X'Y^{(k)}.$$

On dispose maintenant de B estimations du paramètre θ , notées

$$\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \dots, \hat{\theta}^{(B)}$$

Pour donner un ordre d'idée, une valeur de 1000 pour B est couramment utilisée.

A partir de là, on peut

- bâtir un intervalle de confiance pour chacune des composantes de θ , en utilisant la procédure décrite précédemment.
- bâtir un intervalle de confiance pour une prévision $\hat{y}_0 = X_0 \hat{\theta}$: à partir des B estimations bootstrap de θ , on peut construire B prévisions bootstrap

$$\hat{y}_0^{(1)}, \hat{y}_0^{(2)}, \dots, \hat{y}_0^{(B)}$$

et en déduire ainsi un intervalle de confiance.