

Etude de cas

Application en sciences humaines et sociales

Plan du cours

- Bootstrap
- Médiation



Bootstrap

Chapitre 1



Introduction

Objectifs

On cherche à estimer les propriétés d'un estimateur ou d'un prédicteur

- Biais, Variance
- Intervalle de confiance, Test.

Soit X est n -échantillon iid suivant la loi qui admet pour fonction de répartition F .

On estime le paramètre θ par l'estimateur $\hat{\theta}_n = \tau_n(X_1, \dots, X_n) = \tau_n(X)$

La loi de $\hat{\theta}_n - \theta$ dépend de F qui est inconnu.

Question : Comment approcher la loi de $\hat{\theta}_n - \theta$, $P(\hat{\theta}_n - \theta < x) = G_{F,n}(x)$?

Méthode de Monte Carlo

On observe un n -échantillon X iid suivant la loi F .

On veut estimer θ un paramètre de la loi F

Par la simulation on peut approcher la loi de $(\hat{\theta}_n - \theta)$ à n fixé.

Procédure de Monte Carlo

Pour $b=1, \dots, B$

(a) On simule un n -échantillon X_b^{mc} iid suivant la loi F

(b) On calcule $\hat{\theta}_n^b = \tau_n(X_b^{mc})$

On obtient un B -échantillon $\hat{\theta}_n^1, \dots, \hat{\theta}_n^B$ iid suivant la loi de $\hat{\theta}_n = \tau_n(X)$

En pratique F est inconnue

Modèle régulier

Estimateur du MV

Dans un modèle régulier, le maximum de vraisemblance est asymptotiquement efficace : $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{loi} N(0, I^{-1}(\theta))$

De plus si I est continue alors $\sqrt{n}I^{1/2}(\hat{\theta}_n)(\hat{\theta}_n - \theta) \xrightarrow{loi} N(0,1)$

On obtient une **approximation gaussienne** de la loi de $(\hat{\theta}_n - \theta)$:

$$P(\hat{\theta}_n - \theta < x) \approx \Phi\left(x\sqrt{nI(\hat{\theta}_n)}\right) \text{ où } \Phi \text{ est la fct répartition de la loi } N(0,1)$$

→ Intervalle de confiance de niveau asymptotique

→ test asymptotique sur le paramètre θ

Difficultés

1. Forme explicite de l'information de Fisher
2. Qualité de la vitesse de convergence.
3. Modèle non régulier

- Approximation de la loi de $\hat{\theta}_n$ par le processus empirique (ou l'estimateur à noyau de la densité) de l'échantillon $\hat{\theta}_n^1, \dots, \hat{\theta}_n^B$

$$\frac{1}{B} \sum_{i=1}^B I_{]-\infty, x]}(\hat{\theta}_n^i) \xrightarrow[B \rightarrow \infty, n \text{ fixé}]{ps} P(\hat{\theta}_n < x)$$

- Espérance

$$\frac{1}{B} \sum_{i=1}^B \hat{\theta}_n^i \xrightarrow[B \rightarrow \infty, n \text{ fixé}]{ps} E(\hat{\theta}_n)$$

- La variance de l'estimateur en prenant la variance empirique de l'échantillon

$$\frac{1}{B} \sum_{i=1}^B \hat{\theta}_n^{i2} - \left(\frac{1}{B} \sum_{i=1}^B \hat{\theta}_n^i \right)^2 \xrightarrow[B \rightarrow \infty, n \text{ fixé}]{ps} var(\hat{\theta}_n)$$

- Les quantiles de la loi de $\hat{\theta}_n$ par les quantiles empiriques de l'échantillon $(\hat{\theta}_n^1, \dots, \hat{\theta}_n^B)$

On a donc des informations non asymptotiques sur l'estimateur.

On choisit B en fonction de la précision souhaitée

Bootstrap paramétrique

Bootstrap paramétrique

Principe

Modèles paramétrique : on suppose que $X = (X_1, \dots, X_n)$ est un n-échantillon de variables aléatoires iid suivant $F_\theta, \theta \in \Theta$.

On considère $\hat{\theta}_n = \tau_n(X)$ un estimateur de θ .

Idee du Bootstrap paramétrique :

on approche la fonction de répartition de F_θ par $F_{\hat{\theta}_n}$

Définition du bootstrap paramétrique

◦ On cherche à approcher

$$P(\hat{\theta}_n - \theta < x) = G_{\theta,n}(x)$$

* On introduit un n-échantillon $X^* = (X_1^*, \dots, X_n^*)$ de variables aléatoires iid suivant la loi $F_{\hat{\theta}_n}$

Attention $F_{\hat{\theta}_n}$ dépend de l'échantillon X

C'est la loi conditionnelle de X_1^* sachant X qui admet pour fonction de répartition $F_{\hat{\theta}_n}$

* Soit $\hat{\theta}_n^* = \tau_n(X^*)$ l'estimateur calculé sur l'échantillon X^*

« $\hat{\theta}_n^*$ est un estimateur de $\hat{\theta}_n$ »

* On a $P(\hat{\theta}_n^* - \hat{\theta}_n < x | X) = G_{\hat{\theta}_n,n}(x)$

Cette loi conditionnelle de $\hat{\theta}_n^* - \hat{\theta}_n$ sachant X est connue

Échantillon	paramètre	Estimateur
X	θ	$\hat{\theta}_n = \tau_n(X)$
X*	$\hat{\theta}_n$	$\hat{\theta}_n^* = \tau_n(X^*)$

Approximation par bootstrap paramétrique

Si $\hat{\theta}_n$ est un estimateur consistant et $\theta \mapsto G_{\theta,n}(x)$ est continue uniformément en n + des hypothèses de régularité.

alors $|G_{\hat{\theta}_n,n}(x) - G_{\theta,n}(x)| \xrightarrow[n \rightarrow \infty]{proba} 0$ à x fixé.

Autrement dit on peut approcher la loi de $\hat{\theta}_n - \theta$ par la loi conditionnelle de $\hat{\theta}_n^* - \hat{\theta}_n$

$$G_{\theta,n}(x) \approx G_{\hat{\theta}_n,n}(x) = P(\hat{\theta}_n^* - \hat{\theta}_n < x | X)$$

Approximation du biais de $\hat{\theta}_n$: $E(\hat{\theta}_n^* - \hat{\theta}_n | X) = E(\hat{\theta}_n^* | X) - \hat{\theta}_n$

Approximation de la variance de $\hat{\theta}_n$: $var(\hat{\theta}_n^* - \hat{\theta}_n | X) = var(\hat{\theta}_n^* | X)$

Intervalle de confiance & bootstrap paramétrique

- On note $Q_{\hat{\theta}_n}(\alpha)$ le quantile de la loi de $\hat{\theta}_n - \theta$
- $Q_n^*(\alpha) = Q_{\hat{\theta}_n}(\alpha)$ est le quantile de la loi conditionnelle de $\hat{\theta}_n^* - \hat{\theta}_n$ sachant X

$$P(\hat{\theta}_n - \theta < Q_n^*(\alpha)) \approx P(\hat{\theta}_n^* - \hat{\theta}_n < Q_n^*(\alpha) | X) = 1 - \alpha$$

- Intervalle de confiance de niveau $1 - 2\alpha$ pour le paramètre θ :

Par bootstrap paramétrique, il est approché par

$$[\hat{\theta}_n - Q_n^*(1 - \alpha) ; \hat{\theta}_n - Q_n^*(\alpha)]$$

Cet intervalle n'est plus de niveau exact $1 - 2\alpha$ mais il est asymptotique de niveau $1 - 2\alpha$.

- On introduit le niveau exact

$$\beta_n^* = P\left(\theta \in [\hat{\theta}_n - Q_n^*(1 - \alpha) ; \hat{\theta}_n - Q_n^*(\alpha)]\right)$$

Comparaison TCL & bootstrap paramétrique

- On suppose que $\sqrt{n} \frac{1}{\sigma(\theta)} (\hat{\theta}_n - \theta) \xrightarrow{loi} N(0,1) \rightsquigarrow P(\hat{\theta}_n - \theta < x) \approx \Phi\left(\frac{x\sqrt{n}}{\sigma(\hat{\theta}_n)}\right)$
- Approximation de l'IC : $\left[\hat{\theta}_n - \frac{\sigma(\hat{\theta}_n)q_{1-\alpha}}{\sqrt{n}} ; \hat{\theta}_n + \frac{\sigma(\hat{\theta}_n)q_{1-\alpha}}{\sqrt{n}} \right]$
où q_α est le quantile d'ordre α de la loi $N(0,1)$
- On définit le niveau exact de cet intervalle : $\beta_n^{TCL} = P\left(\theta \in [\hat{\theta}_n - \frac{\sigma(\hat{\theta}_n)q_{1-\alpha}}{\sqrt{n}} ; \hat{\theta}_n + \frac{\sigma(\hat{\theta}_n)q_{1-\alpha}}{\sqrt{n}}]\right)$
- Le bootstrap fournit une meilleure approximation que le TCL si

$$|\beta_n^* - (1 - \alpha)| \leq |\beta_n^{TCL} - (1 - \alpha)|$$

En pratique

On utilise une méthode de Monte Carlo pour approcher numériquement les lois conditionnelles de $\hat{\theta}_n^*$ et $\hat{\theta}_n^* - \hat{\theta}_n$.

Algorithme : Pour $b=1, \dots, B$

- On simule un n -échantillon X^{*b} iid suivant la $F_{\hat{\theta}_n}$
- On calcule $\hat{\theta}_n^{*b} = \tau_n(X^{*b})$

$\hat{\theta}_n^{*1}, \dots, \hat{\theta}_n^{*B}$ est un B -échantillon de variables aléatoires iid suivant la loi conditionnelle de $\hat{\theta}_n^*$ sachant X

et $\hat{\theta}_n^{*1} - \hat{\theta}_n, \dots, \hat{\theta}_n^{*B} - \hat{\theta}_n$ suivant celle de $\hat{\theta}_n^* - \hat{\theta}_n$ sachant X

Approximation

- Approximation de l'estimation du biais de $\hat{\theta}_n$ par le bootstrap paramétrique :

$$\frac{1}{B} \sum_{i=1}^B \hat{\theta}_n^{*i} - \hat{\theta}_n$$

- Approximation de l'estimation de la variance de $\hat{\theta}_n$ par le bootstrap paramétrique :

$$\frac{1}{B} \sum_{i=1}^B \hat{\theta}_n^{*i2} - \left(\frac{1}{B} \sum_{i=1}^B \hat{\theta}_n^{*i} \right)^2$$

- Approximation de $Q_n^*(\alpha)$:

le quantile empirique $\hat{Q}_{n,B}^*(\alpha)$ de l'échantillon $(\hat{\theta}_n^{*1} - \hat{\theta}_n, \dots, \hat{\theta}_n^{*B} - \hat{\theta}_n)$

Ces approximations convergent vers les quantités d'intérêt quand $B \rightarrow \infty$ à n fixé

On peut choisir B aussi grand qu'on se souhaite.

Conclusion : deux niveaux d'approximation

- Estimation par bootstrap de la loi $\hat{\theta}_n - \theta$:

$$P(\hat{\theta}_n^* - \hat{\theta}_n < x | X) - P(\hat{\theta}_n - \theta < x) \xrightarrow[n \rightarrow \infty]{} 0$$

- Approximation à partir de l'échantillon simulé $\hat{\theta}_n^{*1} - \hat{\theta}_n, \dots, \hat{\theta}_n^{*B} - \hat{\theta}_n$ de la probabilité conditionnelle $P(\hat{\theta}_n^* - \hat{\theta}_n < x | X)$

$$\frac{1}{B} \sum_{i=1}^B I_{]-\infty, x]}(\hat{\theta}_n^{*i} - \hat{\theta}_n) \xrightarrow[B \rightarrow \infty, n \text{ fixé}]{} P(\hat{\theta}_n^* - \hat{\theta}_n < x | X)$$

- Par exemple pour les quantiles on obtient

$$Q_n^*(\alpha) - Q_{\theta, n}(\alpha) \xrightarrow[n \rightarrow \infty]{} 0$$

$$\hat{Q}_{n, B}^*(\alpha) \xrightarrow[B \rightarrow \infty, n \text{ fixé}]{} Q_n^*(\alpha)$$

Rappel sur le processus empirique

Processus empirique

Définition

Soit $(X_1, \dots, X_n)$ un n-échantillon de variables aléatoires iid suivant F (fonction de répartition de la loi commune)

Définition : Le processus empirique est une fonction aléatoire définie par

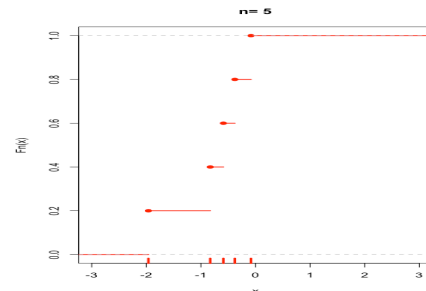
$$F_n(t, \omega) = F_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{]-\infty, t]}(X_i(\omega))$$

Interprétation :

$F_n(t)$ représente la proportion des observations inférieures à t $\rightarrow F_n$ est fct croissante

$$F_n(t) = \begin{cases} 0 & \text{si } t < \min(X_1, \dots, X_n) \\ 1 & \text{si } t \geq \max(X_1, \dots, X_n) \end{cases}$$

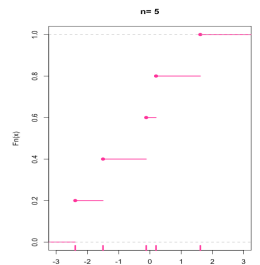
C'est une fonction constante par morceaux càdlàg



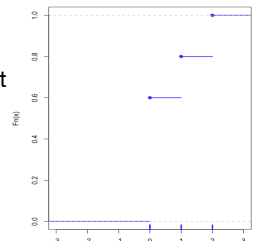
Loi empirique

Le processus empirique F_n est la fonction de répartition d'une loi discrète de support les valeurs prises par $\{X_1(\omega), \dots, X_n(\omega)\}$

- I. Si $\{X_1(\omega), \dots, X_n(\omega)\}$ sont tous distincts, F_n est la fct de répartition de la loi uniforme sur $\{X_1(\omega), \dots, X_n(\omega)\}$



- II. Sinon on note $\{\tilde{X}_1(\omega), \dots, \tilde{X}_m(\omega)\}$ l'échantillon sans répétition et $\hat{p}_j = \frac{1}{n} \text{card}\{i : X_i = \tilde{X}_j\}$ $j = 1, \dots, m$. F_n est la fct de répartition de la loi discrète de support $\{\tilde{X}_1(\omega), \dots, \tilde{X}_m(\omega)\}$ et de probabilité $\{\hat{p}_1, \dots, \hat{p}_m\}$



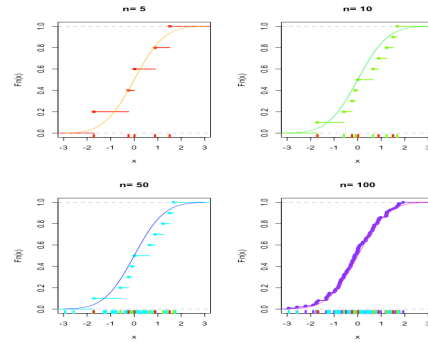
Convergence du processus empirique

Théorème de Glivenko Cantelli :

Si $(X_1, \dots, X_n)$ est un n-échantillon de variables aléatoires iid suivant F

alors

$$\sup_{t \in R} |F_n(t) - F(t)| \xrightarrow[p.s]{n \rightarrow \infty} 0$$



Autres propriétés

Conséquence du Théorème de Glivenko Catelli :

$\forall t \in R, F_n(t)$ est un estimateur fortement consistant de $F(t)$

Conséquence du Théorème Central Limite : $\sqrt{n}(F_n(t) - F(t)) \xrightarrow[loi]{n \rightarrow \infty} \mathcal{N}(0, \sigma_t^2)$

donc $\forall t \in R, F_n(t)$ est un estimateur $\sqrt{n}$ -consistant de $F(t)$

Conséquence du Théorème de Slutsky :

Si $F(t) \neq 0$ et $F(t) \neq 1$ alors $\frac{\sqrt{n}(F_n(t) - F(t))}{\sqrt{F_n(t)(1 - F_n(t))}} \xrightarrow[loi]{n \rightarrow \infty} \mathcal{N}(0, 1)$

Bootstrap non paramétrique

Bootstrap Non paramétrique

Contexte

On suppose que $X = (X_1, \dots, X_n)$ est un n-échantillon de variables aléatoires iid suivant F .

On veut estimer un paramètre de la loi F : $\theta = \theta(F)$

Dans le contexte non paramétrique on peut construire un estimateur à partir de la loi empirique F_n

On estime le paramètre $\theta = \theta(F)$ par $\hat{\theta}_n = \theta(F_n) = \tau_n(X)$

Par exemple

Si $\theta(F) = E(X_1)$ alors $\tau_n(X) = \theta(F_n) = \bar{X}_n$

Si $\theta(F) = F^{-}(1/2)$ (la médiane) alors $\tau_n(X) = \theta(F_n)$ est la médiane de l'échantillon $F_n^{-}(1/2) = \inf\{x : F_n(x) \geq 1/2\}$

Définition du bootstrap non paramétrique

- On veut approcher la loi de $\hat{\theta}_n - \theta$:

$$G_n^F(x) = P(\hat{\theta}_n - \theta < x)$$

Échantillon	Loi	Paramètre	Estimateur
$\mathbf{x}$	F	$\theta(F)$	$\hat{\theta}_n = \tau_n(X) = \theta(F_n)$
$\mathbf{x}^*$	$F_n \equiv \text{Unif}\{X_1, \dots, X_n\}$	$\theta(F_n) = \hat{\theta}_n$	$\hat{\theta}_n^* = \tau_n(X^*)$

- On considère introduit un n-échantillon $X^* = (X_1^*, \dots, X_n^*)$ iid suivant la loi empirique F_n . On l'appelle l'échantillon bootstrappé

Autrement dit conditionnellement à X, les variables $X_1^*, \dots, X_n^*$ sont iid suivant la Loi uniforme sur l'ensemble $\{X_1, \dots, X_n\}$

- Soit $\hat{\theta}_n^* = \tau_n(X^*)$ est l'estimateur calculé sur l'échantillon X^*

Approximation par bootstrap non paramétrique

On approche la loi de $\hat{\theta}_n - \theta$ par la loi conditionnelle de $\hat{\theta}_n^* - \hat{\theta}_n$ sachant X

$$G_n^F(x) \approx G_n^*(x) = P(\hat{\theta}_n^* - \hat{\theta}_n \leq x | X)$$

Approximation du biais de $\hat{\theta}_n$: $E(\hat{\theta}_n^* - \hat{\theta}_n | X) = E(\hat{\theta}_n^* | X) - \hat{\theta}_n$

Approximation de la variance de $\hat{\theta}_n$: $\text{var}(\hat{\theta}_n^* - \hat{\theta}_n | X) = \text{var}(\hat{\theta}_n^* | X)$

Approximation du quantile $Q_n^F(\alpha)$ le quantile de la loi de $\hat{\theta}_n - \theta$ par le quantile

$$Q_n^*(\alpha) = (G_n^*)^{-1}(\alpha) = \inf\{u : G_n^*(u) \geq \alpha\}$$

Loi de $\hat{\theta}_n^* - \hat{\theta}_n$

- La loi conditionnelle de X_1^* sachant X est la loi uniforme sur $\{X_1, \dots, X_n\}$ et conditionnellement à X les va $X_1^*, \dots, X_n^*$ sont iid

- Pour pour $x \in \{X_1, \dots, X_n\}^n$, on a

$$P(X^* = x | X) = P(X_1^* = x_1, \dots, X_n^* = x_n | X)$$

$$= P(X_1^* = x_1 | X) \dots P(X_n^* = x_n | X) = \frac{1}{n^n}$$

- On obtient la loi conditionnelle de $\hat{\theta}_n^* = \tau_n(X^*)$ sachant X

$$\begin{aligned} H_n^*(y) = P(\hat{\theta}_n^* \leq y | X) &= \sum_{x \in \{X_1, \dots, X_n\}^n \text{ tq } \tau_n(x) \leq y} P(X^* = x | X) = \sum_{x \in \{X_1, \dots, X_n\}^n \text{ tq } \tau_n(x) \leq y} \frac{1}{n^n} \\ &= \frac{1}{n^n} \text{card}\{x \in \{X_1, \dots, X_n\}^n \text{ tq } \tau_n(x) \leq y\} \end{aligned}$$

- Puis on en déduit la loi conditionnelle de $\hat{\theta}_n^* - \hat{\theta}_n$ sachant X

$$G_n^*(y) = P(\hat{\theta}_n^* - \hat{\theta}_n \leq y | X) = H_n(y + \hat{\theta}_n)$$

Exemple : estimation de la moyenne

$$\theta(F) = E(X_1)$$

- On suppose que $X_1 \in L^2$

- L'estimateur de $E(X_1)$ est $\hat{\theta}_n = \theta(F_n) = \bar{X}_n$

- C'est un estimateur sans biais de variance $\text{var}(\hat{\theta}) = \frac{1}{n} \text{var}(X_1)$

- On a $\hat{\theta}_n^* = \bar{X}_n^*$, c'est la moyenne empirique de l'échantillon bootstrappé X^*

- Estimation du biais : $E(\bar{X}_n^* - \bar{X}_n | X) = \frac{1}{n} \sum_{i=1}^n E(X_i^* | X) - \bar{X}_n = 0$ car $E(X_i^* | X) = \bar{X}_n$

- Estimation de la variance : $\text{var}(\bar{X}_n^* - \bar{X}_n | X) = \text{var}(\bar{X}_n^* | X) = \frac{1}{n} \text{var}(X_1^* | X) = \frac{1}{n^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \frac{1}{n} \text{var}(X_1) = \text{var}(\hat{\theta}_n)$

Justification de l'approximation par bootstrap pour la moyenne

Si $X_1, \dots, X_n$ sont iid dans L^2 alors

1. TCL classique : $\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{loi} N(0, \sigma^2)$
2. (Admis) De plus la loi conditionnelle de $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ converge en loi vers une loi gaussienne $N(0, \sigma^2)$

Autrement dit pour tout x on a

$$P\left(\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \theta) < x\right) \xrightarrow{n \rightarrow \infty} \Phi(x) \text{ et } P\left(\frac{\sqrt{n}}{\sigma}(\bar{X}_n^* - \bar{X}_n) < x | X\right) \xrightarrow{n \rightarrow \infty} \Phi(x)$$

D'où

$$P\left(\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \theta) < x\right) \sim P\left(\frac{\sqrt{n}}{\sigma}(\bar{X}_n^* - \bar{X}_n) < x | X\right) \text{ quand } n \rightarrow \infty$$

Approximation

1. Approximation de l'estimation du biais de $\hat{\theta}_n$ par le bootstrap paramétrique :

$$\frac{1}{B} \sum_{i=1}^B \hat{\theta}_n^{*i} - \hat{\theta}_n$$

2. Approximation de l'estimation de la variance de $\hat{\theta}_n$ par le bootstrap paramétrique :

$$\frac{1}{B} \sum_{i=1}^B \hat{\theta}_n^{*i2} - \left(\frac{1}{B} \sum_{i=1}^B \hat{\theta}_n^{*i} \right)^2$$

3. Approximation de $Q_n^*(\alpha)$:

le quantile empirique $\hat{Q}_{n,B}^*(\alpha)$ de l'échantillon $(\hat{\theta}_n^{*1} - \hat{\theta}_n, \dots, \hat{\theta}_n^{*B} - \hat{\theta}_n)$

Ces approximations convergent vers les quantités d'intérêt quand $B \rightarrow \infty$ à n fixé

En pratique on peut choisir B aussi grand qu'on se souhaite.

En pratique

On approche la loi de $\hat{\theta}_n^* - \hat{\theta}_n$ par la simulation (méthode de Monte Carlo)

On adapte la méthode proposée dans le cas du bootstrap paramétrique en remplaçant $F_{\hat{\theta}_n}$ par F_n

Algorithme : Pour $b=1, \dots, B$

- (a) On simule un n -échantillon X^{*b} iid suivant la loi uniforme sur $\{X_1, \dots, X_n\}$
- (b) On calcule $\hat{\theta}_n^{*b} = \tau_n(X^{*b})$

$\hat{\theta}_n^{*1}, \dots, \hat{\theta}_n^{*B}$ est un B -échantillon de $\hat{\theta}_n^* - \hat{\theta}_n$ iid simulé suivant la loi conditionnelle de $\hat{\theta}_n^*$ sachant X

Intervalle de confiance par bootstrap

$Q_n^F(\alpha)$ le quantile d'ordre α de la loi de $\hat{\theta}_n - \theta$

IC d'ordre $1 - 2\alpha$: $IC_{1-2\alpha}(\theta) = [\hat{\theta}_n - Q_n^F(1 - \alpha); \hat{\theta}_n - Q_n^F(\alpha)]$

Approximation par bootstrap $IC_{1-2\alpha}^*(\theta) = [\hat{\theta}_n - Q_n^*(1 - \alpha); \hat{\theta}_n - Q_n^*(\alpha)]$

où $Q_n^*(\alpha)$ est le quantile d'ordre α de la loi conditionnelle de $\hat{\theta}_n^* - \hat{\theta}_n$ sachant X

En pratique

Approximation par bootstrap avec simulation $\widehat{IC}^*(\theta) = [\hat{\theta}_n - \hat{Q}_{n,B}^*(1 - \alpha); \hat{\theta}_n - \hat{Q}_{n,B}^*(\alpha)]$

où $\hat{Q}_n^*(\alpha)$ est le quantile empirique d'ordre α de l'échantillon simulé $(\hat{\theta}_n^{*1} - \hat{\theta}_n, \dots, \hat{\theta}_n^{*B} - \hat{\theta}_n)$

Intervalle de confiance naïf

Un intervalle de confiance naïf aurait pu être

$$IC_{1-2\alpha}^{naïf}(\theta) = [J_n^*(\alpha) ; J_n^*(1 - \alpha)]$$

où $J_n^*(\alpha)$ est le quantile d'ordre α de la loi conditionnelle de $\hat{\theta}_n^*$ sachant X

En pratique

Approximation par bootstrap avec simulation $\widehat{IC}^{naïf}(\theta) = [\hat{J}_n^*(\alpha) ; \hat{J}_n^*(1 - \alpha)]$

où $\hat{J}_n^*(\alpha)$ est le quantile empirique d'ordre α de l'échantillon simulé $(\hat{\theta}_n^{*1}, \dots, \hat{\theta}_n^{*B})$

Double Bootstrap

On se place dans le contexte où la variance $\sigma^2(\theta)$ est difficile à obtenir et/ou estimer

Le bootstrap fournit une estimation de la loi de $\sqrt{n}(\hat{\theta}_n - \theta)$ par $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$

Etape 1 : on approche la loi de $\sqrt{n}(\hat{\theta}_n - \theta)$ par celle de $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ avec $\hat{\theta}_n = \tau_n(X)$ et $\hat{\theta}_n^* = \tau_n(X^*)$ et $X^* = (X_1^*, \dots, X_n^*)$ iid $Unif\{X_1, \dots, X_n\}$

Etape 2 : on estime l'écart type de $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ en utilisant $\sqrt{n}(\hat{\theta}_n^{**} - \hat{\theta}_n^*)$

Avec

$$\hat{\theta}_n^{**} = \tau_n(X^{**})$$

$X^{**} = (X_1^{**}, \dots, X_n^{**})$ iid suivant la loi uniforme sur $\{X_1^*, \dots, X_n^*\}$

Bootstrap-t : intervalle de confiance studentisé

On construit souvent des estimateurs tels que $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{loi} N(0, \sigma^2(\theta))$

En estimant la variance de la loi limite, on peut studentiser $\sqrt{n}(\hat{\theta}_n - \theta)$

C'est à dire $T_n = \frac{\sqrt{n}}{\sigma(\hat{\theta}_n)}(\hat{\theta}_n - \theta) \xrightarrow{loi} N(0, 1)$ [T_n est une fonction asymptotiquement pivotale]

$$IC_{1-2\alpha}^T(\theta) = \left[\hat{\theta}_n - \frac{\sigma(\hat{\theta}_n)q_{1-\alpha}}{\sqrt{n}} ; \hat{\theta}_n + \frac{\sigma(\hat{\theta}_n)q_{1-\alpha}}{\sqrt{n}} \right] \text{ où } q_\alpha \text{ est le quantile d'ordre } \alpha \text{ de la loi } N(0,1)$$

En pratique

Approximation par bootstrap avec simulation : A partir de $\hat{\theta}_n^{*1}, \dots, \hat{\theta}_n^{*B}$ simulés suivant la loi conditionnelle de $\hat{\theta}_n^*$

sachant X on construit $t_n^{*i} = \frac{\sqrt{n}}{\sigma(\hat{\theta}_n^{*i})}(\hat{\theta}_n^{*i} - \hat{\theta}_n)$ pour $i=1, \dots, B$

On calcule $\hat{J}_n^{*}(\alpha)$ le quantile empirique d'ordre α de l'échantillon $(t_n^{*1}, \dots, t_n^{*B})$

$$\widehat{IC}^{T^*}(\theta) = \left[\hat{\theta}_n - \frac{\sigma(\hat{\theta}_n)\hat{J}_n^*(1-\alpha)}{\sqrt{n}} ; \hat{\theta}_n - \frac{\sigma(\hat{\theta}_n)\hat{J}_n^*(\alpha)}{\sqrt{n}} \right]$$

Algorithme

Double bootstrap

- Pour $d=1, \dots, B$
 - On simule $X^{*d} = (X_1^{*d}, \dots, X_n^{*d})$ iid suivant la loi uniforme sur $\{X_1, \dots, X_n\}$
 - On calcule $\hat{\theta}_n^{*d} = \tau_n(X^{*d})$
 - Pour $b=1, \dots, B$
 - On simule $X^{**b} = (X_1^{**b}, \dots, X_n^{**b})$ iid suivant la loi uniforme sur $\{X_1^{*d}, \dots, X_n^{*d}\}$
 - On calcule $N_b^d = \tau_n(X^{**b}) - \hat{\theta}_n^{*d} = \hat{\theta}_n^{**b} - \hat{\theta}_n^{*d}$
 - On calcule $\hat{\sigma}_n^d$ l'écart type empirique de l'échantillon $N_1^d, \dots, N_B^d$
 - On calcule $W_d^* = \frac{\theta_n^{*d} - \hat{\theta}_n}{\hat{\sigma}_n^d}$

Intervalle de confiance

Double bootstrap

- On calcule $\hat{\omega}_n^*(\alpha)$ le quantile empirique d'ordre α de l'échantillon simulé $(W^{*1}, \dots, W^{*B})$ avec $W_d^* = \frac{\theta_n^{*d} - \hat{\theta}_n}{\hat{\sigma}_n^d}$ $d=1, \dots, B$

- On pose $\bar{\sigma} = \frac{1}{B} \sum_{d=1}^B \sigma_n^d$ la moyenne empirique de l'échantillon $\hat{\sigma}_n^1, \dots, \hat{\sigma}_n^B$

- L'intervalle de confiance est

$$\widehat{IC}^{double*}(\theta) = \left[\hat{\theta}_n - \bar{\sigma} \hat{\omega}_n^*(1 - \alpha) ; \hat{\theta}_n - \bar{\sigma} \hat{\omega}_n^*(\alpha) \right]$$

Comparaison des intervalles

Pour comparer les différents intervalles de confiance approchés

On compare les probabilités de couverture exactes avec $1 - 2\alpha$

La meilleure approche est celle pour laquelle la probabilité de couverture est la plus proche de $1 - 2\alpha$

La longueur de l'intervalle est un critère de sélection **seulement** si les intervalles de confiance ont le même niveau exact.

Modèle uniforme

Un contre exemple au bootstrap non paramétrique

- On suppose que $X = (X_1, \dots, X_n)$ sont iid suivant la loi uniforme sur $[0, \theta]$ et on estime θ par le maximum de l'échantillon $M_n = \max(X_1, \dots, X_n)$
- On veut approcher par bootstrap non paramétrique

$$G_n(x) = P(M_n - \theta \leq x) = \left(1 + \frac{x}{\theta}\right)^n 1_{[-\theta, 0]}(x) + 1_{R^+}(x)$$

- On construit $(X_1^*, \dots, X_n^*)$ iid suivant la loi uniforme sur $X_1, \dots, X_n$ (conditionnellement à X) et on note $M_n^* = \max(X_1^*, \dots, X_n^*)$ le maximum de l'échantillon bootstrappé
- On a $P(M_n^* - M_n \geq 0 | X_1, \dots, X_n) = 1 - (1 - 1/n)^n \rightarrow 1 - e^{-1}$
- L'estimation par bootstrap non paramétrique n'est donc pas consistante

Modèle uniforme

bootstrap paramétrique

- On construit $(X_1^{**}, \dots, X_n^{**})$ iid suivant la loi uniforme sur $[0, M_n]$ (conditionnellement à X) et on note $M_n^{**} = \max(X_1^{**}, \dots, X_n^{**})$ le maximum de l'échantillon bootstrappé
- On a $P(M_n^{**} - M_n \leq x | X) = \left(1 + \frac{x}{M_n}\right)^n 1_{[-M_n, 0]}(x) + 1_{R^+}(x)$
- L'estimation par bootstrap paramétrique est consistante

Bilan

Le principe du bootstrap repose sur le fait que l'on peut approcher la loi du n-échantillon X

La méthode d'estimation peut être facilement mise en oeuvre en utilisant la simulation et les méthodes de Monte Carlo

Pour des variables aléatoires iid, il suffit donc d'estimer la loi marginale

Bootstrap NP : estimation de F_{X_1} par le processus empirique.

Bootstrap P : estimation de F_{X_1} par $F_{\hat{\theta}_n}$

Problèmes :

Comment étendre la méthode aux observations dépendantes et / ou non identiquement distribuées ?

Exemples : Séries temporelles, la régression ...

Modèle semi paramétrique

Modèle ARMA

On suppose $(X_t)_{t \in \mathbb{Z}}$ est un modèle ARMA(p,q) causal inversible

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q}$$

avec $w_t \sim \text{wn}(0, \sigma_w^2)$

On estime les coefficients AR et MA par une méthode de vraisemblance approchée.

On note $\hat{\phi}, \hat{\theta}$ les estimateurs des coefficients AR et MA

Pour évaluer les intervalles de confiance, on utilise l'approximation asymptotique gaussienne de la loi des estimateurs

Les limites sont obtenues sous l'hypothèse que le bruit est une suite de va iid (bruit blanc fort)

On ne fait pas d'hypothèse paramétrique sur la loi du bruit

Bootstrap et séries temporelles

Bootstrap sur les résidus

Sous l'hypothèse que les résidus sont iid, on peut appliquer le principe du bootstrap non paramétrique sur les résidus.

Les résidus ne sont pas observés, mais on peut appliquer le bootstrap sur les résidus estimés.

Principe :

1. A partir du modèle estimé on approche les résidus $\hat{w}_1, \dots, \hat{w}_n$
2. On construit un échantillon bootstrappé du bruit $\hat{w}_1^*, \dots, \hat{w}_n^*$.
Conditionnellement à $X_1, \dots, X_n$, les variables $\hat{w}_1^*, \dots, \hat{w}_n^*$ sont iid suivant la loi uniforme sur $\{\hat{w}_1, \dots, \hat{w}_n\}$
3. On en déduit un échantillon bootstrappé du processus X
$$\begin{cases} X_t^* = X_t & \text{si } t < \max(p, q) \\ X_t^* = \hat{\phi}_1 X_{t-1}^* + \dots + \hat{\phi}_p X_{t-p}^* + w_t^* + \hat{\theta}_1 w_{t-1}^* + \dots + \hat{\theta}_q w_{t-q}^* & \text{si } t \geq \max(p, q) \end{cases}$$
4. On calcule les estimateurs de ϕ et θ sur $X_1^*, \dots, X_n^*$. On note les estimateurs ϕ^*, θ^*

Approximation

- On approche la loi de $\begin{pmatrix} \hat{\phi} \\ \hat{\theta} \end{pmatrix} - \begin{pmatrix} \phi \\ \theta \end{pmatrix}$ par la loi conditionnelle de $\begin{pmatrix} \phi^* \\ \theta^* \end{pmatrix} - \begin{pmatrix} \hat{\phi} \\ \hat{\theta} \end{pmatrix}$ sachant X qui est connue
- Cette loi peut être facilement approchée par la simulation
- On pose $\lambda = \begin{pmatrix} \phi \\ \theta \end{pmatrix}$
- Pour chacun des coefficients $\lambda_i, i = 1, \dots, p + q$ on construit un intervalle de confiance de niveau asymptotique $1 - \alpha$ en prenant

$$\lambda_i \in [\hat{\lambda}_i - q_i^*(1 - \alpha/2), \hat{\lambda}_i - q_i^*(\alpha/2)]$$
où $q_i^*(\alpha)$ est le quantile d'ordre α de la loi de $\lambda_i^* - \hat{\lambda}_i$

En pratique

On approche la loi de $\begin{pmatrix} \phi^* \\ \theta^* \end{pmatrix} - \begin{pmatrix} \hat{\phi} \\ \hat{\theta} \end{pmatrix}$ par la simulation

- Pour $b=1, \dots, B$
 - On simule un n-échantillon $(\hat{w}_1^{*b}, \dots, \hat{w}_n^{*b})$ iid suivant la loi uniforme sur $\{\hat{w}_1, \dots, \hat{w}_n\}$
 - On calcule $X_1^{*b}, \dots, X_n^{*b}$
 - On calcule les estimateurs ϕ^{*b}, θ^{*b}
- On approche les quantiles $q_i^*(\alpha)$ par les quantiles empiriques des B-échantillons (coordonnée par coordonnée) de $\begin{pmatrix} \phi^{*1} \\ \theta^{*1} \end{pmatrix} - \begin{pmatrix} \hat{\phi} \\ \hat{\theta} \end{pmatrix} \dots \dots \begin{pmatrix} \phi^{*B} \\ \theta^{*B} \end{pmatrix} - \begin{pmatrix} \hat{\phi} \\ \hat{\theta} \end{pmatrix}$

Prévision et Bootstrap

On approche la loi de $\hat{X}_{n+h} - X_{n+h}$ par la loi de $\hat{X}_{n+h}^* - \hat{X}_{n+h}$

où $X_{1+n}^*, \dots, X_{n+h}^*$ est la prévision calculée à partir de $X_1^*, \dots, X_n^*$ et des coefficients estimés ϕ^*, θ^*

En pratique on ajoute une étape à l'algorithme précédent :

- Pour $b=1, \dots, B$ l'étape additionnelle est
 - On prévoit à l'horizon h $(X_{1+n}^{*b}, \dots, X_{n+h}^{*b})$ à partir de $X_1^*, \dots, X_n^*$ et ϕ^{*b}, θ^{*b}
- Pour $j=1, \dots, h$ on calcule $\hat{x}_{n+j}^*(\alpha)$ le quantile empirique d'ordre α du B-échantillon

$$(X_{n+j}^{*1} - \hat{X}_{n+j}, \dots, X_{n+j}^{*B} - \hat{X}_{n+j})$$
- On en déduit un intervalle de niveau asymptotique $1 - \alpha$ en prenant
- $X_{n+j} \in [\hat{X}_{n+h} - \hat{x}_{n+j}^*(1 - \alpha/2); \hat{X}_{n+h} - \hat{x}_{n+j}^*(\alpha/2)]$

Situation paramétrique

On suppose $(X_t)_{t \in \mathbb{Z}}$ est un modèle ARMA(p,q) causal inversible

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q}$$

avec un bruit iid centré L^2 et on suppose que $w_t \sim g_\nu$

On estime ϕ, θ et le paramètre supplémentaire ν

On note les estimateurs $\hat{\phi}, \hat{\theta}, \hat{\nu}$

Principe

1. On construit un échantillon bootstrappé du bruit $\hat{w}_1^*, \dots, \hat{w}_n^*$.
Conditionnellement à $X_1, \dots, X_n$, les variables $\hat{w}_1^*, \dots, \hat{w}_n^*$ sont iid suivant la loi de densité $g_{\hat{w}}$
2. On en déduit un échantillon bootstrappé du processus X

$$\begin{cases} X_t^* = X_t & \text{si } t < \max(p, q) \\ X_t^* = \hat{\phi}_1 X_{t-1}^* + \dots + \hat{\phi}_p X_{t-p}^* + w_t^* + \hat{\theta}_1 w_{t-1}^* + \dots + \hat{\theta}_q w_{t-q}^* & \text{si } t \geq \max(p, q) \end{cases}$$
3. On calcule les estimateurs de ϕ et θ sur $X_1^*, \dots, X_n^*$. On note les estimateurs ϕ^*, θ^*
4. On calcule l'estimateur de ν sur $X_1^*, \dots, X_n^*$ noté ν^*

Approche par bloc

On observe $X_1, \dots, X_n$.

On construit $N = n - \ell + 1$ blocs de longueur ℓ

$$\begin{aligned} \mathbb{B}_1 &= (X_1, X_2, \dots, X_\ell), \\ \mathbb{B}_2 &= (X_2, \dots, X_\ell, X_{\ell+1}), \\ &\dots \\ \mathbb{B}_N &= (X_{n-\ell+1}, \dots, X_n), \end{aligned}$$

Soit $b = n/\ell$ le nombre de blocs pour reconstruire un échantillon de longueur

(on suppose que b est un entier)

On considère $i_1^*, \dots, i_b^*$ iid suivant la loi uniforme sur $\{1, \dots, N\}$

Le processus bootstrappé est obtenu par concaténation des blocs

$$X^* = [\mathbb{B}_{i_1^*}, \dots, \mathbb{B}_{i_b^*}]$$

Situation non paramétrique

Hypothèse : $(X_t)_{t \in \mathbb{Z}}$ est un processus stationnaire

Remarque

- En prenant $\ell = 1$ on retrouve le bootstrap classique non paramétrique qui est adapté uniquement au cas indépendant stationnaire (iid)
- Pour valider cette approche pour des variables dépendantes, il faut prendre une longueur de bloc $\ell = \ell_n$ tel que

$$\ell_n^{-1} + n^{-1} \ell_n \rightarrow 0 \quad (n \rightarrow \infty)$$

- En pratique on choisit la longueur en fonction de nombre d'autocorrélations empiriques $\hat{\rho}_n(h)$ non nulles
- On choisit ℓ_n tel que $\hat{\rho}_n(\ell_n)$ est la dernière autocorrélation non nulle.

Forme de bloc alternative

- Une alternative est de prendre des blocs disjoints
- On construit b blocs de longueur ℓ en supposant que $b = \frac{n}{\ell}$ est un entier

$$\begin{aligned}\tilde{\mathbb{B}}_1 &= (X_1, \dots, X_\ell), \\ \tilde{\mathbb{B}}_2 &= (X_{\ell+1}, \dots, X_{2\ell}), \\ &\dots \\ \tilde{\mathbb{B}}_b &= (X_{(b-1)\ell+1}, \dots, X_n).\end{aligned}$$

- On considère $i_1^*, \dots, i_b^*$ iid suivant la loi uniforme sur $\{1, \dots, b\}$
- Le processus bootstrappé est obtenu par concaténation des blocs

$$X^* = [\tilde{\mathbb{B}}_{i_1^*}, \dots, \tilde{\mathbb{B}}_{i_b^*}]$$

Bootstrap stationnaire

C'est une version de bootstrap par bloc mais on ajoute de l'aléatoire dans la construction des blocs

On suppose que

- la longueur du bloc
- le début du bloc

sont des variables aléatoires

Avec cette construction aléatoire des blocs la série bootstrappée est aussi stationnaire

Cette propriété n'est pas satisfaite par les méthodes par blocs déterministes

Le principe

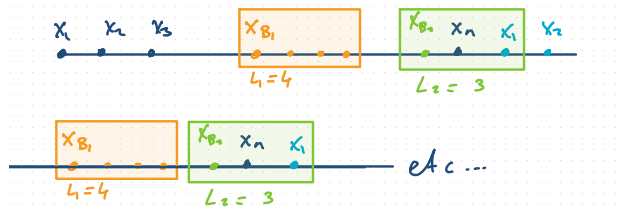
1. On simule les longueurs de bloc $L_1, \dots, L_N$ iid avec un nombre de bloc N est aléatoire $N = \inf\{k : \sum_{i=1}^k L_i \geq n\}$
2. On simule les débuts de bloc $B_1, \dots, B_N$ iid suivant la loi uniforme sur $\{1, \dots, n\}$

$$X^* = (X_{B_1} \dots X_{B_1+L_1-1} X_{B_2} \dots X_{B_2+L_2-1} \dots X_{B_N} \dots X_{B_N+L_N-1})$$

Remarque

Le dernier bloc $X_{B_N} \dots X_{B_N+L_N-1}$ est éventuellement tronqué pour obtenir une série bootstrappée de longueur n

On prolonge la série $X_1, \dots, X_n$ de façon cyclique $X_1, \dots, X_n X_1, \dots, X_n$ si $B_j + L_j - 1 > n$



Remarque

Le choix classique de la distribution sur la longueur est la loi géométrique de paramètre p

Rappel si Z suit une loi géométrique : $P(Z = k) = p^* (1 - p)^{k-1}$ avec $k \in \mathbb{N}^*$

En théorie : il faut prendre $p = p_n$ tel que $p_n \rightarrow 0$ et $np_n \rightarrow \infty$

En pratique : on choisit p_n tel que la longueur moyenne coïncide avec le choix effectué pour les blocs déterministes

$$\ell_n = p_n^{-1}$$

Médiation

Chapitre 2



L'analyse de médiation est une méthode statistique utilisée pour évaluer la façon dont une variable X transmet son effet sur une variable conséquent Y.

Quel est le mécanisme par lequel X influence Y ?

Modèle de régression

Régression

Les modèles de régression permettent d'expliquer une variable Y à partir de X

$$Y = f(X) + \epsilon$$

où ϵ est bruit blanc centré.

La fonction f contient de l'information sur la façon dont X influence Y.

Si f est constant alors X n'a pas d'effet sur Y.

Objectif : Mesurer/quantifier l'effet de X sur Y .

Peut on évaluer cet effet à partir de f uniquement ?

Modèle de régression linéaire

On explique l'influence de X sur Y par la relation

$$Y = \beta X + \alpha + \epsilon$$

Interprétation du coefficient β : il mesure l'effet de X sur Y.

Pour tester H_0 : « X n'influence pas Y » (autrement dit l'absence d'un effet de X sur Y)

On teste si le coefficient β est significatif $H_0 : \beta = 0$

Méthodologie :

1. Test de student
2. Application du bootstrap

Régression $Y = X\beta + \epsilon$: bootstrap sur les couples

Si les X sont supposés aléatoires, alors on applique le bootstrap sur les couples :

$$(X_i, Y_i) \in R^{p+1} \text{ avec } i=1, \dots, n.$$

Le principe :

1. Soit $I_1, \dots, I_n$ iid suivant la loi uniforme sur $\{1, \dots, n\}$
2. On pose $(X_i^*, Y_i^*) = (X_{I_i}, Y_{I_i})$ pour $i=1, \dots, n$
3. On calcule l'estimateur des MCO : $\hat{\beta}^* = (X^{*T} X^*)^{-1} X^{*T} Y^*$

Régression $Y = X\beta + \epsilon$: bootstrap sur les résidus

Lorsque les variables X sont déterministes, on applique le bootstrap sur les résidus

Le principe :

1. On estime le paramètre β par les MCO $\hat{\beta} = (X^T X)^{-1} X^T Y$
2. On estime les résidus $\hat{\epsilon} = Y - X\hat{\beta} = \begin{pmatrix} \hat{\epsilon}_1 \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix}$
3. On construit $\hat{\epsilon}^* = (\hat{\epsilon}_1^* \dots \hat{\epsilon}_n^*)$ iid suivant la loi uniforme sur $(\hat{\epsilon}_1 \dots \hat{\epsilon}_n)$ et $Y^* = X\hat{\beta} + \epsilon^*$
4. On calcule l'estimateur des MCO sur (X, Y^*) : $\hat{\beta}^* = (X^T X)^{-1} X^T Y^*$

Bootstrap et régression

On approche la loi de $\hat{\beta} - \beta$ par la loi de $\hat{\beta}^* - \hat{\beta}$

Soit $q_i^*(\alpha)$ le quantile d'ordre α de $\hat{\beta}_i^* - \hat{\beta}_i$ conditionnellement à (X, Y)

L'intervalle de confiance de niveau $1 - 2\alpha$ approché par bootstrap est

$$\beta_i \in [\hat{\beta}_i - q_i^*(1 - \alpha), \hat{\beta}_i - q_i^*(\alpha)]$$

Test de significatif du coefficient β_i :

Région critique du test de niveau α : On rejette l'hypothèse nulle $\beta_i = 0$ si

$$0 \notin [\hat{\beta}_i - q_i^*(1 - \alpha/2), \hat{\beta}_i - q_i^*(\alpha/2)]$$

Limite de cette approche pour évaluer l'effet.

On considère deux variables (X,Y). Quel est l'effet de X sur Y ?

On construit un modèle de régression.

$a = \text{lm}(Y \sim X)$

Coefficients:

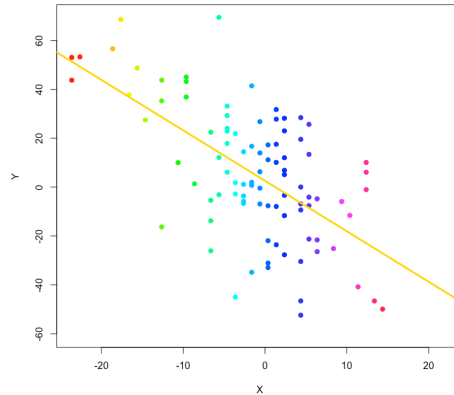
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.6028	1.3399	1.196	0.235
X	-1.7060	0.1314	-12.982	<2e-16 ***

Conclusion :

l'effet de X sur Y est significatif

De plus

$\beta \in [-1.7 \pm 0.26]$ (95%)



Ajout d'une variable explicative

On ajoute une variable explicative M dans la construction du modèle de régression.

$b = \text{lm}(Y \sim X + M)$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.42409	1.03156	2.350	0.0208 *
X	-0.08843	0.21816	-0.405	0.6961
M	-0.87390	0.10455	-8.359	4.62e-13 ***

Le coefficient de X n'est pas significatif

Au vu des résultats de ce modèle X n'a pas d'effet sur Y.

Ceci contredit le résultat précédent.

Conclusion :

X n'influence pas directement Y mais X influence Y à travers M

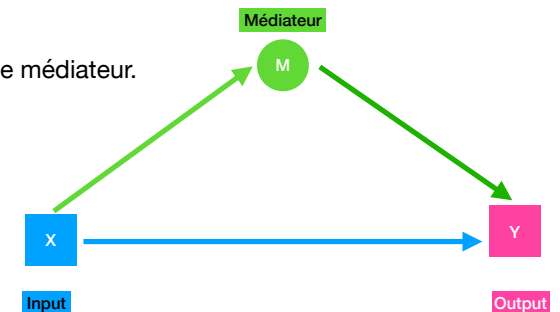
Médiation

Définition

Un modèle de médiation est un système causal, « X influence Y », dans lequel la variable X influence partiellement ou totalement la valeur de Y par le biais d'une variable M.

Définition : La variable M est appelée médiateur.

Le médiateur est donc une variable qui permet d'expliquer le processus par lequel X influence Y.



Exemple : la persuasion et le changement d'attitude

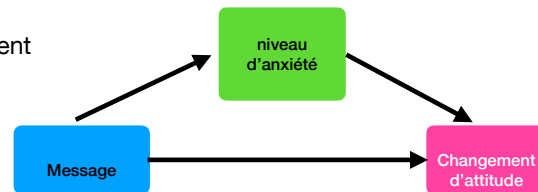
Est-ce que le fait de formuler un message en termes de gains plutôt qu'en termes de pertes (X) influence le comportement?

Campagne anti tabac :

Est-ce que le fait de formuler un message antitabac en termes de gains plutôt qu'en termes de pertes (X) influence la probabilité d'un arrêt du tabac (Y) ?

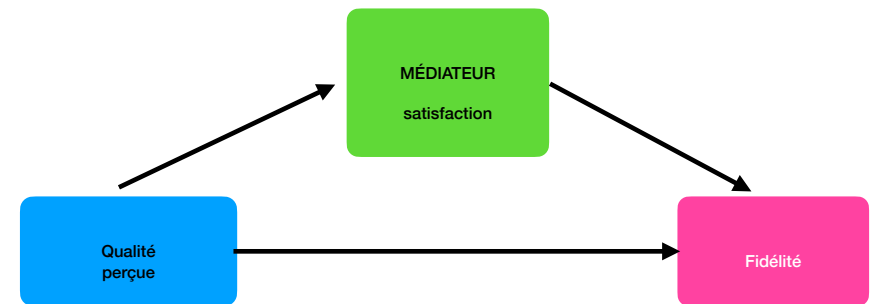
Le type de message influence le niveau d'anxiété des gens, qui à son tour influence le comportement ?

Le niveau d'anxiété est potentiellement un médiateur de l'effet du message sur la proba d'arrêter de fumer.



Exemple en marketing

Comment la qualité perçue influence la fidélité ?

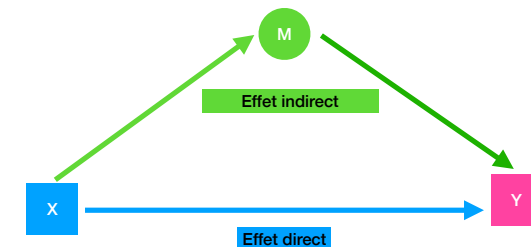
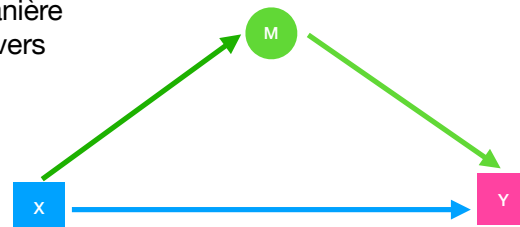


Effets dans le modèle de médiation

- Il existe deux chemins par lesquels X peut influencer Y.
- L'effet direct de X sur Y : le chemin de X à Y sans passer par M
- L'effet indirect de X sur Y à travers M : le chemin de X à Y passant par M.

- L'effet indirect représente la manière dont Y est influencé par X à travers une séquence causale :

X influence M, qui à son tour influence Y.



Comment mesurer les effets de X sur Y, direct et indirect via M ?

Comment tester si les effets direct et indirect sont significatifs ?

Modèle linéaire de médiation

Définition des effets

- **Effet total :**
l'effet total de X sur Y est défini à partir du modèle $Y = i + \beta X + \epsilon$
Il est égal au paramètre β

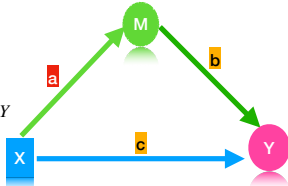
En présence d'un médiateur cet effet se décompose comme
la somme de l'effet direct et de l'effet indirect à travers M

- **Effet direct :**
L'effet direct est donné par le coefficient c dans l'équation $Y = i_Y + cX + bM + \epsilon_Y$.

- **Effet indirect :**
- $\begin{cases} Y = i_Y + cX + bM + \epsilon_Y \\ M = i_M + aX + \epsilon_M \end{cases} \Rightarrow Y = i_Y + cX + b(i_M + aX + \epsilon_M) + \epsilon_Y$
 $\Rightarrow Y = i_Y + bi_M + (c + ab)X + b\epsilon_M + \epsilon_Y$

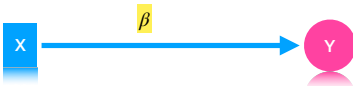
On a donc $\beta = c + ab$

L'effet indirect est donc égal au produit $ab = \beta - c$



Modèle

On suppose que X influence Y : $Y = i + \beta X + \epsilon$

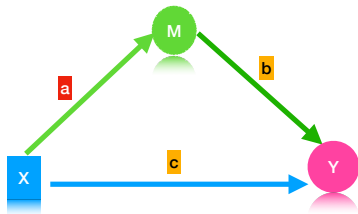


Si X à un effet sur Y via un médiateur M,

On décrit les relations entre X,M,Y par le système d'équations linéaires :

$$\begin{cases} Y = i_Y + cX + bM + \epsilon_Y \\ M = i_M + aX + \epsilon_M \end{cases}$$

Ce système est le modèle de médiation



X is déterministe
Y and M sont aléatoires

Dans le modèle linéaire de médiation

$$\begin{cases} Y = i_Y + cX + bM + \epsilon_Y \\ M = i_M + aX + \epsilon_M \end{cases}$$

On peut donc évaluer l'effet de X sur Y à partir des coefficients du modèle

Effet Direct	c
Effet Indirect	ab
Effet Total	c+ab

Estimation des paramètres

On estime les paramètres par la méthode des moindres carrés appliquée sur les deux équations du système linéaire.

$$\begin{cases} Y = i_Y + cX + bM + \epsilon_Y \\ M = i_M + aX + \epsilon_M \end{cases}$$

On pose $\theta_Y = (i_Y, c, b)$

L'estimateur des MCO est $\hat{\theta}_Y = ({}^t\mathbb{X}_1\mathbb{X}_1)^{-1} {}^t\mathbb{X}_1Y$ où $\mathbb{X}_1 = [1, X, M]$

On pose $\theta_M = (i_M, a)$

L'estimateur des MCO est $\hat{\theta}_M = ({}^t\mathbb{X}_2\mathbb{X}_2)^{-1} {}^t\mathbb{X}_2M$ où $\mathbb{X}_2 = [1, X]$

Estimation de l'effet indirect

L'effet indirect est égal au produit ab

Un estimateur de l'effet indirect est donné par le produit des estimateurs des MCO : $\hat{a}\hat{b}$

Objectifs :

1. Construction d'un intervalle de confiance
2. Tester l'existence d'un effet indirect

Il faut donc obtenir une approximation de loi de $\hat{a}\hat{b} - ab$

Estimation de l'effet direct

Comme l'effet direct est un des coefficients du modèle de régression

$$Y = i_Y + cX + bM + \epsilon_Y$$

L'estimation est immédiate c'est l'estimateur des MCO du paramètre c

Le test de significativité du coefficient c (test de Student) permet de tester l'existence d'un effet direct de X sur Y

$$H_0 : c = 0 \text{ contre } c \neq 0$$

Approximation par la Δ -méthode

On suppose que les bruits ϵ_Y, ϵ_M sont des bruits blancs gaussiens

La vraisemblance s'écrit

$$\begin{aligned} V(\theta_Y, \theta_M, \sigma_M^2, \sigma_Y^2) &= f_{\theta_Y, \theta_M, \sigma_M^2, \sigma_Y^2}(Y, M | X) \\ &= f_{\theta_Y, \theta_M, \sigma_M^2, \sigma_Y^2}(Y | M, X) f_{\theta_Y, \theta_M}(M | X) = f_{\theta_Y, \sigma_Y^2}(Y | M, X) f_{\theta_M, \sigma_M^2}(M | X) \end{aligned}$$

D'où

$$\log(V(\theta_Y, \theta_M, \sigma_M^2, \sigma_Y^2)) = \log(f_{\theta_Y, \sigma_Y^2}(Y | M, X)) + \log(f_{\theta_M, \sigma_M^2}(M | X))$$

Calculer l'estimateur du maximum de vraisemblance des paramètres $(\theta_Y, \theta_M, \sigma_M^2, \sigma_Y^2)$ est équivalent à calculer

$$(\theta_Y, \sigma_Y^2) \text{ l'estimateur du MV (= MCO) du modèle } Y = i_Y + cX + bM + \epsilon_Y$$

$$(\theta_M, \sigma_M^2) \text{ l'estimateur du MV (= MCO) du modèle } M = i_M + aX + \epsilon_M$$

Approximation par bootstrap sur les résidus

A partir des paramètres estimés par la méthode des MCO, on calcule les résidus estimés :

$$\hat{\epsilon}_1^Y, \dots, \hat{\epsilon}_n^Y \text{ et } \hat{\epsilon}_1^M, \dots, \hat{\epsilon}_n^M$$

Application du bootstrap sur les résidus :

1. On simule $\hat{\epsilon}_1^{M*}, \dots, \hat{\epsilon}_n^{M*}$ iid suivant la loi uniforme sur $(\hat{\epsilon}_1^M, \dots, \hat{\epsilon}_n^M)$
2. On calcule $M_j^* = \hat{i}_M + \hat{a}X_j + \epsilon_j^{M*}$ pour $j=1, \dots, n$
3. On simule $\hat{\epsilon}_1^{Y*}, \dots, \hat{\epsilon}_n^{Y*}$ iid suivant la loi uniforme sur $(\hat{\epsilon}_1^Y, \dots, \hat{\epsilon}_n^Y)$
4. On calcule $Y_j^* = \hat{i}_Y + \hat{c}X_j + \hat{b}M_j^* + \epsilon_j^{Y*}$ pour $j=1, \dots, n$
5. On calcule les estimateurs des MCO sur l'échantillon bootstrappé et l'estimation de l'effet indirect $\hat{a}^*\hat{b}^*$

Comme le modèle est régulier, on a la normalité asymptotique

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\text{loi}} N(0, \Sigma_\theta) \text{ où } \Sigma_\theta = \begin{pmatrix} \Sigma_{\theta_Y, \sigma_Y} & 0 \\ 0 & \Sigma_{\theta_M, \sigma_M} \end{pmatrix} \text{ est l'inverse de l'information de Fisher}$$

Les estimateurs des paramètres (θ_M, σ_M^2) et (θ_Y, σ_Y^2) sont asymptotiquement indépendants.

Conséquence :

$$\sqrt{n} \begin{pmatrix} \hat{a} - a \\ \hat{b} - b \end{pmatrix} \xrightarrow{\text{loi}} N \left(0, \begin{pmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{pmatrix} \right)$$

En appliquant la Δ -méthode $(x, y) \mapsto xy$ on obtient

$$\sqrt{n}(\hat{a}\hat{b} - ab) \xrightarrow{\text{loi}} N(0, b^2\sigma_a^2 + a^2\sigma_b^2)$$

On estime la variance de la loi limite, puis on en déduit un intervalle de confiance.

On approche la loi de $\hat{a}\hat{b} - ab$ par la loi de $\hat{a}^*\hat{b}^* - \hat{a}\hat{b}$

On en déduit un intervalle de confiance sur ab en prenant

$$ab \in [\hat{a}\hat{b} - q^*(1 - \alpha/2); \hat{a}\hat{b} - q^*(\alpha/2)]$$

où $q^*(\alpha)$ est le quantile d'ordre α de la loi de $\hat{a}^*\hat{b}^* - \hat{a}\hat{b}$

Test : on teste l'existence de l'effet indirect à partir de cet intervalle

Si $0 \in [\hat{a}\hat{b} - q^*(1 - \alpha/2); \hat{a}\hat{b} - q^*(\alpha/2)]$ on accepte l'absence d'un effet indirect.