ULB

Section 2 : Estimateurs des moindres carrés

STAT-F-406

Master en sciences mathématiques, Master en statistique

ACTU-F4001

Master en sciences actuarielles

Davy Paindaveine

Université libre de Bruxelles

2023–2024

Estimateur de β par analogie

Soient $(Y_1, X_1), \ldots, (Y_n, X_n)$ des observations i.i.d. engendrées depuis le modèle linéaire faible

$$Y = \beta' X + \varepsilon,$$

où $\mathrm{E}[XX'] > 0$ (nous supposerons toujours $\mathrm{E}[XX'] > 0$ dans la suite).

Alors l'estimateur naturel de $\beta = (\mathrm{E}[XX'])^{-1}\mathrm{E}[XY]$ est

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right).$$

Par la LFGN, $\hat{\beta} \xrightarrow{\text{p.s.}} \beta$ si $n \to \infty$.

Attention aux notations: X_j, X_i, X_{ij}.

Cas particulier du modèle de position

Le modèle de position est obtenu avec $k = 1$ et $X = 1$ p.s.:

$$Y = \beta + \varepsilon.$$

Dans ce modèle, l'estimateur de

$$\beta = (\mathrm{E}[XX'])^{-1}\mathrm{E}[XY] = \mathrm{E}[Y]$$

ci-dessus est

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} Y_i \overset{\text{not}}{=} \bar{Y}.$$

Estimateur des moindres carrés ordinaires

Pour chaque valeur de β, on peut considérer les résidus

$$e_i(\beta) = Y_i - \beta' X_i, \quad i = 1, \ldots, n.$$

Intuitivement, β est une bonne valeur du paramètre si les résidus sont petits (voir le slide suivant). Il est classique de minimiser

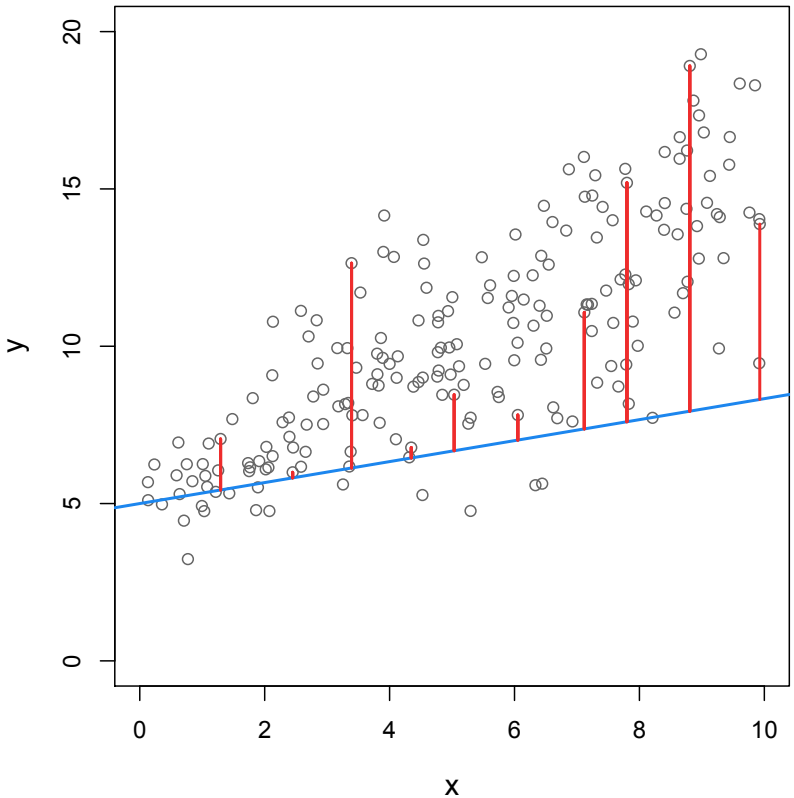
$$\sum_{i=1}^{n} |e_i(\beta)| \quad \text{ou} \quad \sum_{i=1}^{n} (e_i(\beta))^2.$$

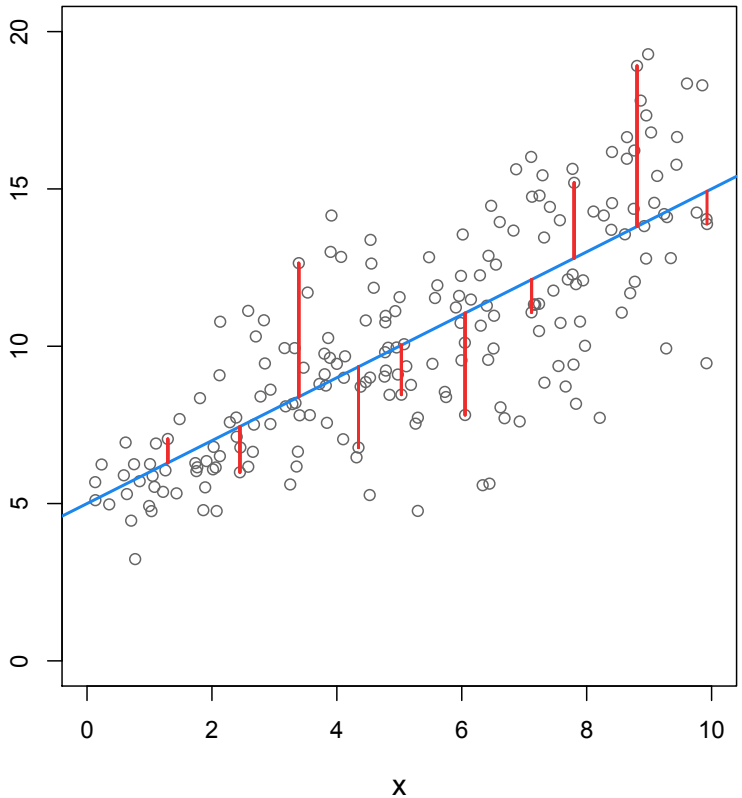
En fait,

$$\hat{\beta}_{\text{MCO}} \stackrel{\text{def}}{=} \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} (e_i(\beta))^2 = \hat{\beta},$$

d'où la terminologie estimateur des moindres carrés (ordinaires) pour $\hat{\beta}$.

Estimateur des moindres carrés ordinaires





Estimateur des moindres carrés ordinaires

Puisque

$$\begin{aligned}
\nabla_\beta \textstyle\sum_{i=1}^n (e_i(\beta))^2 &= \nabla_\beta \textstyle\sum_{i=1}^n (Y_i - \beta' X_i)^2 \\
&= \textstyle\sum_{i=1}^n 2(Y_i - \beta' X_i)(-X_i) \\
&= -2\big\{(\textstyle\sum_{i=1}^n X_i Y_i) - (\textstyle\sum_{i=1}^n X_i \beta' X_i)\big\} \\
&= -2\big\{(\textstyle\sum_{i=1}^n X_i Y_i) - (\textstyle\sum_{i=1}^n X_i X_i')\beta\big\} \\
&= -2n\big\{(\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i Y_i) - (\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i X_i')\beta\big\},
\end{aligned}$$

la seule valeur de β qui annule le gradient de la fonction objectif

$$\beta \mapsto \textstyle\sum_{i=1}^n (e_i(\beta))^2$$

est $\hat{\beta}$. En étudiant la matrice hessienne, on voit que cette fonction objectif est convexe sur $\mathbb{R}^k$, ce qui implique que $\hat{\beta}$ est un minimum global.

Estimateur des moindres carrés ordinaires

L'estimateur

$$\hat{\beta}_{\mathrm{LAD}} \overset{\mathrm{def}}{=} \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} |e_i(\beta)|$$

n'admet pas d'expression explicite, mais il a aussi de bonnes propriétés, notamment en termes de robustesse.

Non-biais

Nous allons étudier les propriétés de l'estimateur $\hat{\beta}$.

Théorème 1

Dans un modèle semi-fort, $\hat{\beta}$ est sans biais pour β.

Preuve: En notant $\mathrm{E}[\cdot|\mathcal{X}]$ l'espérance conditionnelle sachant $X_1, \ldots, X_n$ et $Q = \frac{1}{n}\sum_{i=1}^n X_i X_i'$, on a

$$\mathrm{E}[\hat{\beta}|\mathcal{X}] = \mathrm{E}[Q^{-1}(\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i Y_i)|\mathcal{X}] = Q^{-1}\tfrac{1}{n}\textstyle\sum_{i=1}^n \mathrm{E}[X_i Y_i|\mathcal{X}]$$

$$= Q^{-1}\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i \mathrm{E}[Y_i|\mathcal{X}] = Q^{-1}\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i \mathrm{E}[X_i'\beta + \varepsilon_i|\mathcal{X}]$$

$$= Q^{-1}\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i (\mathrm{E}[X_i'\beta|\mathcal{X}] + \mathrm{E}[\varepsilon_i|\mathcal{X}]) = Q^{-1}\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i (X_i'\beta + 0) = \beta$$

pour tout β. Donc $\mathrm{E}[\hat{\beta}] = \beta$ pour tout β. □

Variance

Si $\hat{\beta}$ vise donc bien en moyenne, il pourrait avoir une grande variance.

Théorème 2

Dans un modèle semi-fort, $\mathrm{Var}[\hat{\beta}|\mathcal{X}] = \frac{1}{n^2} Q^{-1} \sum_{i=1}^{n} \sigma^2(X_i) X_i X_i' Q^{-1}$.

Preuve: En utilisant des notations similaires à la preuve précédente,

$$\mathrm{Var}[\hat{\beta}|\mathcal{X}] = \mathrm{Var}[Q^{-1}(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i Y_i)|\mathcal{X}] = \tfrac{1}{n^2} Q^{-1} \mathrm{Var}[\textstyle\sum_{i=1}^{n} X_i Y_i|\mathcal{X}] Q^{-1}$$

$$= \tfrac{1}{n^2} Q^{-1} \textstyle\sum_{i=1}^{n} \mathrm{Var}[X_i Y_i|\mathcal{X}] Q^{-1} = \tfrac{1}{n^2} Q^{-1} \textstyle\sum_{i=1}^{n} X_i \mathrm{Var}[Y_i|\mathcal{X}] X_i' Q^{-1}$$

$$= \tfrac{1}{n^2} Q^{-1} \textstyle\sum_{i=1}^{n} X_i \mathrm{Var}[\cancel{X_i'\beta} + \varepsilon_i|\mathcal{X}] X_i' Q^{-1} = \tfrac{1}{n^2} Q^{-1} \textstyle\sum_{i=1}^{n} \sigma^2(X_i) X_i X_i' Q^{-1},$$

puisque, par définition, $\sigma^2(X_i) = \mathrm{Var}[\varepsilon_i|X_i]$. □

Variance

Théorème 2

Dans un modèle semi-fort, $\mathrm{Var}[\hat{\beta}|\mathcal{X}] = \frac{1}{n^2} Q^{-1} \sum_{i=1}^{n} \sigma^2(X_i) X_i X_i' Q^{-1}$.

Quelques remarques:

- Dans un modèle semi-fort homoscédastique,

$$\begin{aligned} \mathrm{Var}[\hat{\beta}|\mathcal{X}] &= \frac{1}{n^2} Q^{-1} \sum_{i=1}^{n} \sigma^2 X_i X_i' Q^{-1} \\ &= \frac{\sigma^2}{n} Q^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i'\right) Q^{-1} \\ &= \frac{\sigma^2}{n} Q^{-1}. \end{aligned}$$

- On a $\mathrm{Var}[\hat{\beta}] = \mathrm{E}[\mathrm{Var}[\hat{\beta}|\mathcal{X}]] + \mathrm{Var}[\mathrm{E}[\hat{\beta}|\mathcal{X}]] = \mathrm{E}[\mathrm{Var}[\hat{\beta}|\mathcal{X}]]$.
- Si $n \to \infty$, alors $n \,\mathrm{Var}[\hat{\beta}|\mathcal{X}] \to (\mathrm{E}[XX'])^{-1} \mathrm{E}[\sigma^2(X) XX'] (\mathrm{E}[XX'])^{-1}$, de sorte que la variance conditionnelle tend vers zéro.

Loi exacte

La variance qui tend vers zéro confirme la convergence de $\hat{\beta}$.
Mais pour faire de l'inférence fondée sur $\hat{\beta}$, on a besoin de sa loi.

Théorème 3

Dans le modèle fort avec normalité, $\hat{\beta}|\mathcal{X} \sim \mathcal{N}_k\left(\beta, \frac{\sigma^2}{n} Q^{-1}\right)$.

Preuve: Dans le modèle fort avec normalité, $\varepsilon_i|X_i \sim \mathcal{N}(0, \sigma^2)$. Puisque les observations sont i.i.d., on a donc $\varepsilon_i|\mathcal{X} \sim \mathcal{N}(0, \sigma^2)$, ce qui livre

$$(X_i'\beta + \varepsilon_i)|\mathcal{X} \sim \mathcal{N}(X_i'\beta, \sigma^2), \quad \text{et donc} \quad X_iY_i|\mathcal{X} \sim \mathcal{N}(X_iX_i'\beta, \sigma^2 X_iX_i').$$

Par additivité des normales indépendantes, on a donc

$$\textstyle\sum_{i=1}^n X_iY_i|\mathcal{X} \sim \mathcal{N}(\sum_{i=1}^n X_iX_i'\beta, \sigma^2 \sum_{i=1}^n X_iX_i'),$$

ce qui fournit le résultat (pourquoi?) □

Loi asymptotique

Si on n'est pas dans le modèle restrictif dans lequel on a pu déterminer la loi exacte de $\hat{\beta}$, on peut obtenir sa loi asymptotique.

Théorème 4

Même dans le modèle faible, mais pour autant que $\mathrm{E}[\sigma^2(X)\|X\|^2] < \infty$, *on a que* $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{D}} \mathcal{N}_k\big(0, (\mathrm{E}[XX'])^{-1}\mathrm{E}[\sigma^2(X)XX'](\mathrm{E}[XX'])^{-1}\big)$.

Preuve: Puisque

$$\hat{\beta} = Q^{-1}\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i(X_i'\beta + \varepsilon_i) = \beta + Q^{-1}\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i\varepsilon_i,$$

le lemme de Slutsky et le TCL fournissent

$$\sqrt{n}(\hat{\beta} - \beta) = Q^{-1}\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n \varepsilon_i X_i \xrightarrow{\mathcal{D}} (\mathrm{E}[XX'])^{-1}\mathcal{N}_k\big(0, \mathrm{E}[(\varepsilon X)(\varepsilon X)']\big),$$

ce qui établit le résultat (puisque $\mathrm{E}[\varepsilon^2 XX'] = \mathrm{E}[\sigma^2(X)XX']$). □

Efficacité (1)

Peut-on faire plus précis que $\hat{\beta}$? En tout cas, pas asymptotiquement dans le modèle fort avec normalité, car...

Théorème 5

Dans le modèle fort avec normalité, $\hat{\beta}$ est l'estimateur du maximum de vraisemblance de β.

Preuve: Dans le modèle considéré, la vraisemblance

$$L_{\beta,\sigma^2} = \prod_{i=1}^n f_{\beta,\sigma^2}^{(Y_i,X_i)}(Y_i,X_i) = \prod_{i=1}^n f_{\beta,\sigma^2}^{Y_i|X_i}(Y_i,X_i) f_{\beta,\sigma^2}^{X_i}(X_i)$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(Y_i-\beta' X_i)^2\Big) \prod_{i=1}^n f^{X_i}(X_i)$$

$$= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i-\beta' X_i)^2\Big) \prod_{i=1}^n f^{X_i}(X_i)$$

est maximale en β lorsque $\sum_{i=1}^n (Y_i-\beta' X_i)^2$ est minimale en β. □

Efficacité (2)

Pour n fixé et pour le modèle semi-fort homoscédastique, on a le

Théorème 6 (Gauss–Markov)

Dans le modèle semi-fort homoscédastique, $\hat{\beta}$ est à variance conditionnellement minimale dans la classe $\mathcal{C}$ des estimateurs linéaires et conditionnellement sans biais.

Soit $\tilde{\beta}$ un estimateur de β. Ici,
$\tilde{\beta}$ est linéaire $\overset{\text{def}}{\Leftrightarrow} \tilde{\beta} = \sum_{i=1}^n \tilde{W}_i Y_i$ pour des v.a. $\tilde{W}_i = \tilde{W}_i(X_1, \ldots, X_n)$
$\tilde{\beta}$ est conditionnellement sans biais $\overset{\text{def}}{\Leftrightarrow} \mathrm{E}[\tilde{\beta}|\mathcal{X}] = \beta$ pour tout $\beta \in \mathbb{R}^k$

Puisque $\hat{\beta} = Q^{-1}(\frac{1}{n}\sum_{i=1}^n X_i Y_i) = \sum_{i=1}^n (Q^{-1}\frac{1}{n}X_i)Y_i = \sum_{i=1}^n \hat{W}_i Y_i$,
$\hat{\beta}$ satisfait bien ces deux propriétés. Gauss–Markov affirme qu'il est le BLUE (Best Linear Unbiased Estimator).

Efficacité (2)

Preuve: Soit $\tilde{\beta} \in \mathcal{C}$. Notons d'abord que, pour tout $\beta \in \mathbb{R}^k$,

$$\beta = \mathrm{E}[\tilde{\beta}|\mathcal{X}] = \mathrm{E}[\textstyle\sum_{i=1}^n \tilde{W}_i Y_i|\mathcal{X}] = \sum_{i=1}^n \tilde{W}_i \mathrm{E}[Y_i|\mathcal{X}] = \sum_{i=1}^n \tilde{W}_i X_i' \beta,$$

de sorte que $\sum_{i=1}^n \tilde{W}_i X_i' = I_k$ p.s. (I_k est la matrice identité $k \times k$).

Donc

$$\begin{aligned}
\mathrm{Var}[\tilde{\beta}|\mathcal{X}] &= \mathrm{Var}[\textstyle\sum_{i=1}^n \tilde{W}_i Y_i|\mathcal{X}] = \sum_{i=1}^n \mathrm{Var}[\tilde{W}_i Y_i|\mathcal{X}] \\
&= \textstyle\sum_{i=1}^n \tilde{W}_i \mathrm{Var}[Y_i|\mathcal{X}] \tilde{W}_i' = \sum_{i=1}^n \tilde{W}_i \mathrm{Var}[X_i'\beta + \varepsilon_i|\mathcal{X}] \tilde{W}_i' \\
&= \textstyle\sum_{i=1}^n \tilde{W}_i \mathrm{Var}[\varepsilon_i|X_i] \tilde{W}_i' = \sigma^2 \sum_{i=1}^n \tilde{W}_i \tilde{W}_i'.
\end{aligned}$$

En particulier, $\mathrm{Var}[\hat{\beta}|\mathcal{X}] = \sigma^2 \sum_{i=1}^n \hat{W}_i \hat{W}_i'$.

Efficacité (2)

On a donc

$$\begin{aligned}
\mathrm{Var}[\tilde{\beta}|\mathcal{X}] &= \sigma^2 \textstyle\sum_{i=1}^n \tilde{W}_i \tilde{W}_i' \\
&= \sigma^2 \textstyle\sum_{i=1}^n \{\hat{W}_i + (\tilde{W}_i - \hat{W}_i)\}\{\hat{W}_i + (\tilde{W}_i - \hat{W}_i)\}' \\
&= \mathrm{Var}[\hat{\beta}|\mathcal{X}] + T + T' + \sigma^2 \textstyle\sum_{i=1}^n (\tilde{W}_i - \hat{W}_i)(\tilde{W}_i - \hat{W}_i)' \\
&= \mathrm{Var}[\hat{\beta}|\mathcal{X}] + \sigma^2 \textstyle\sum_{i=1}^n (\tilde{W}_i - \hat{W}_i)(\tilde{W}_i - \hat{W}_i)' \\
&\geq \mathrm{Var}[\hat{\beta}|\mathcal{X}]
\end{aligned}$$

(au sens de l'ordre associé aux matrices semi-définies positives), où on a utilisé le fait que

$$\begin{aligned}
T \stackrel{\text{def}}{=} \sigma^2 \textstyle\sum_{i=1}^n (\tilde{W}_i - \hat{W}_i)\hat{W}_i' &= \sigma^2 \textstyle\sum_{i=1}^n (\tilde{W}_i - Q^{-1}\frac{1}{n}X_i)(Q^{-1}\frac{1}{n}X_i)' \\
&= \tfrac{\sigma^2}{n}\big(\textstyle\sum_{i=1}^n \tilde{W}_i X_i' - Q^{-1}\frac{1}{n}\sum_{i=1}^n X_i X_i'\big)Q^{-1} = 0. \qquad \square
\end{aligned}$$

Estimateur des moindres carrés pondérés

Dans le modèle semi-fort hétéroscédastique, $\hat{\beta}$ n'est pas efficace. Comment construire un estimateur efficace?

Le modèle

$$\frac{Y}{\sigma(X)} = \Big(\frac{X}{\sigma(X)}\Big)'\beta + \frac{\varepsilon}{\sigma(X)},$$

que nous noterons

$$\tilde{Y} = \tilde{X}'\beta + \tilde{\varepsilon}$$

est semi-fort homoscédastique. En effet,

$$\mathrm{E}[\tilde{\varepsilon}|\tilde{X}] = \mathrm{E}\left[\mathrm{E}\left[\frac{\varepsilon}{\sigma(X)}|X\right]|\tilde{X}\right] = \mathrm{E}\left[\frac{1}{\sigma(X)}\mathrm{E}[\varepsilon|X]|\tilde{X}\right] = 0$$

$$\mathrm{Var}[\tilde{\varepsilon}|\tilde{X}] = \mathrm{E}[\tilde{\varepsilon}^2|\tilde{X}] = \mathrm{E}\left[\mathrm{E}\left[\frac{\varepsilon^2}{\sigma^2(X)}|X\right]|\tilde{X}\right] = \mathrm{E}\left[\frac{1}{\sigma^2(X)}\mathrm{E}[\varepsilon^2|X]|\tilde{X}\right] = 1.$$

Estimateur des moindres carrés pondérés

Donc l'estimateur des moindres carrés pondérés

$$\begin{aligned}\hat{\beta}_{\mathrm{MCP}} &= \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{Y}_i\right)\\ &= \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma^2(X_i)}X_iX_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma^2(X_i)}X_iY_i\right)\end{aligned}$$

est efficace pour β.

Malheureusement, il ne s'agit pas d'un estimateur...

Comment dès lors implémenter cet estimateur?

Estimateur des moindres carrés pondérés

Ceci requiert typiquement de choisir un modèle pour $x \mapsto \sigma^2(x)$.
Par exemple, $\sigma^2(x) = a + x'Bx$, avec $a > 0$ et B semi-définie positive.

Procédure:

(1) Calculer les résidus $e_i = Y_i - X_i'\hat{\beta}$, $i = 1, \ldots, n$, fondés sur $\hat{\beta}$

(2) Estimer a et B par moindres carrés ordinaires dans le modèle linéaire

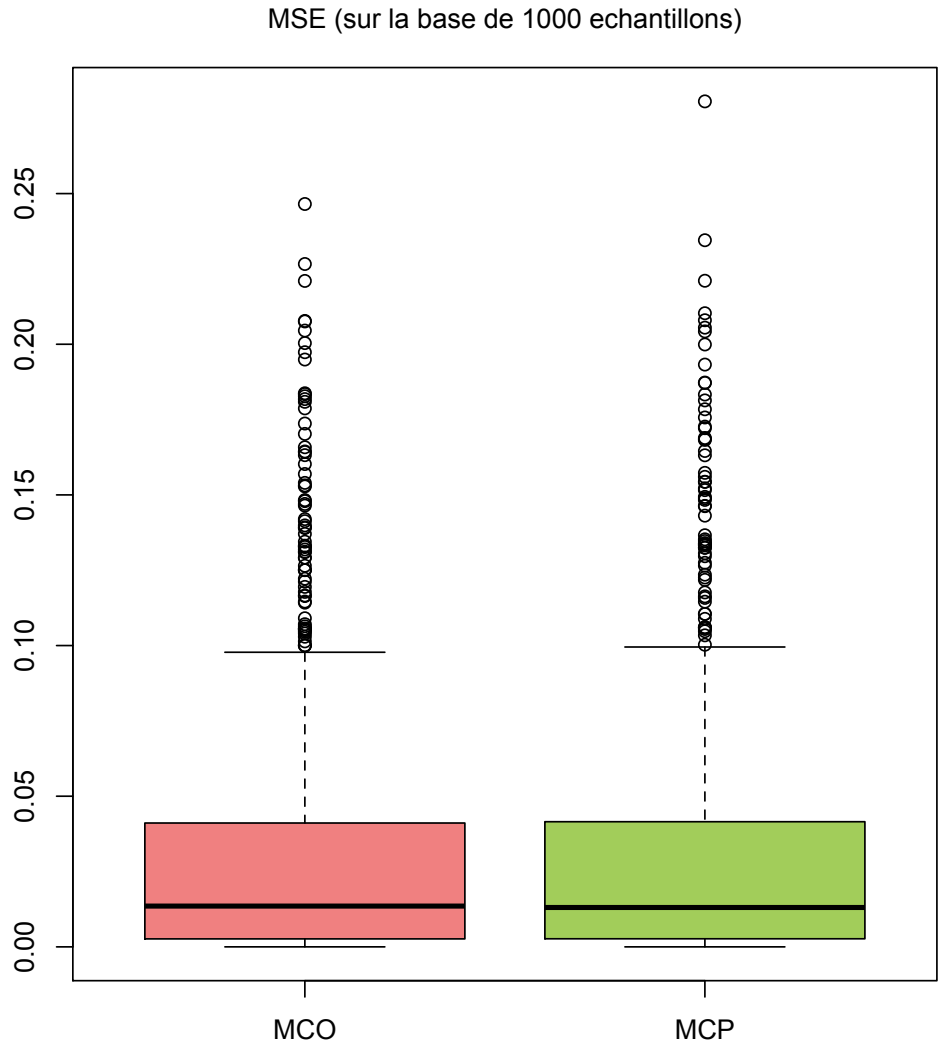
$$e_i^2 = a + X_i'BX_i + \eta_i$$

(convergence car $\varepsilon^2 = a + X'BX + \eta$ est un modèle linéaire semi-fort)

(3) Calculer $\hat{\beta}_{\mathrm{MCP}}$, fondé sur $\hat{\sigma}^2(x) = \hat{a} + x'\hat{B}x$.

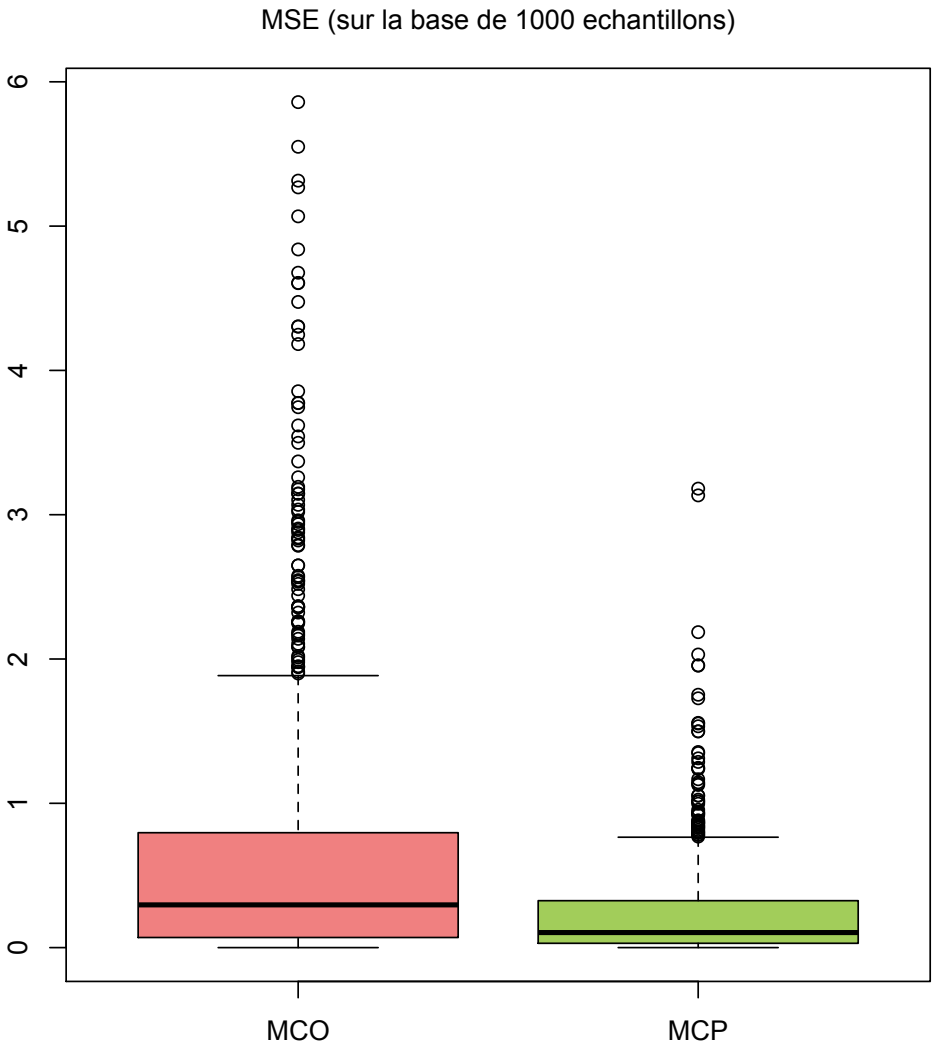
On peut montrer que l'estimateur $\hat{\beta}_{\mathrm{MCP}}$ qui en résulte est (asymptotiquement) efficace malgré les estimations non efficaces aux étapes (1)–(2).

Estimateur des moindres carrés pondérés



$$Y = 5 + 2X + \varepsilon$$
$$X \sim \text{Unif}(1, 10)$$
$$\varepsilon | X \sim \mathcal{N}(0, 1)$$
$$n = 200$$

Estimateur des moindres carrés pondérés



$$Y = 5 + 2X + \varepsilon$$
$$X \sim \text{Unif}(1, 10)$$
$$\varepsilon | X \sim \mathcal{N}(0, 1+X^2)$$
$$n = 200$$