Les Moindres Carrés ordinaires

Stéphane Adjemian

Université Maine, GAINS & CEPREMAP

2 mars 2008

– Le processus générateur des données (DGP) est de la forme suivante :

$$y_i = \mu + \varepsilon_i$$

pour $i = 1, \ldots, N$, avec $\mu \in \mathbb{R}$ un paramètre inconnu et $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, 1)$.

– Ainsi, la variable aléatoire y_i est normalement distribuée d'espérance μ et de variance 1.

– On cherche à estimer l'espérance μ.

– La forme du DGP suggère de régresser $\{y_i\}_{i=1}^N$ sur une constante.

– Le modèle estimé est donc de la forme suivante :

$$y_i = a + \epsilon_i$$

pour $i = 1, \ldots, N$.

– Par définition, on construit l'estimateur des MCO de a en minimisant la somme des carrés des erreurs :

$$\hat{a}_N = \arg\min_{\{a\}} \sum_{i=1}^{N} (y_i - a)^2$$

– La CNO associée à ce programme est :

$$2 \sum_{i=1}^{N} (y_i - \hat{a}_N) = 0$$

– De façon équivalente nous avons :

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \hat{a}_N$$

En notant que sur le membre de droite nous sommons N fois $\hat{a}_N$ (qui ne dépend pas de i) il vient finalement :

$$\hat{a}_N = N^{-1} \sum_{i=1}^{N} y_i \equiv \bar{y}_N$$

– L'estimateur de la constante est simplement la moyenne arithmétique des $\{y_i\}_{i=1}^{N}$.

– Nous devons maintenant établir le rapport entre $\hat{a}_N$ et μ...

– $\hat{a}_N$ est un estimateur sans biais de μ ssi $\mathbb{E}[\hat{a}_N] = \mu$.
L'estimateur $\hat{a}_N$ est sans biais ssi en moyenne – pour différents échantillons – il est égale à μ.

– Par définition de l'estimateur nous avons :

$$\mathbb{E}[\hat{a}_N] = \mathbb{E}\left[N^{-1}\sum_{i=1}^{N} y_i\right] = N^{-1}\sum_{i=1}^{N}\mathbb{E}[y_i]$$

où la deuxième égalité vient de la linéarité de l'espérance.

– Or, d'aprés le DGP, nous savons que $\mathbb{E}[y_i] = \mu$ pour tout i.
Nous avons donc $\mathbb{E}[\hat{a}_N] = N^{-1} \times N \times \mu = \mu$. **L'estimateur est sans biais.**

– Quelle est la précision de cet estimateur ? Rien ne sert d'avoir un estimateur sans biais s'il n'est pas précis...

– Calculons la variance de $\hat{a}_N$:

$$\mathbb{V}\left[\hat{a}_N\right] = \mathbb{V}\left[N^{-1}\sum_{i=1}^{N} y_i\right] = N^{-2}\mathbb{V}\left[\sum_{i=1}^{N} y_i\right]$$

– Or, d'après le DGP, y_i est indépendant de y_j pour tout $j \neq i$. Nous avons donc :

$$\mathbb{V}\left[\hat{a}_N\right] = N^{-2}\sum_{i=1}^{N} \mathbb{V}\left[y_i\right] = \frac{1}{N}$$

car $\mathbb{V}[y_i] = 1$ pour tout i.

– $\hat{a}_N$ est un estimateur d'autant plus précis de μ que la taille de l'échantillon est importante.

– On note que si N tend vers l'infini alors la variance de $\hat{a}_N$ tend vers 0 (la précision tend vers l'infini). Ceci nous assure que $\hat{a}_N$ tend vers μ en probabilité lorsque N tend vers l'infini, c'est-à-dire que :

$$\lim_{N \to \infty} \mathbb{P}\left(|\hat{a}_N - \mu| \geq \nu\right) = 0.$$

pour tout $\nu > 0$ arbitrairement petit.

– Au total, estimer l'espérance de la variable aléatoire y à l'aide d'une moyenne arithmétique semble être une bonne idée.

– Soit le DGP suivant :

$$y_i = a + bx_i + \epsilon_i$$

avec a et b deux paramètres réels, ϵ_i une variable aléatoire identiquement et indépendamment distribuée vérifiant :

(i) $\mathbb{E}\big[\epsilon_i|\mathbf{x}\big] = 0$, et

(ii) $\mathbb{E}\big[\epsilon_i^2|\mathbf{x}\big] = \sigma^2$ pour tout i.

– Notons que les propriétés de la perturbation ϵ_i sont définies conditionellement à la variable explicative $\mathbf{x} \equiv (x_1, \ldots, x_N)$. On peut néanmoins en déduire des propriétés pour les moments non conditionnels de ϵ_i. On utilise le **théorème des espérances itérées.**

– Soit (X, Y) un couple de variables aléatoires réelles. On note $f_{X,Y}(x, y)$ la densité jointe associée au couple de v.a.r.

– On obtient la densité marginale de X en intégrant la densité jointe par rapport à y :

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)\mathrm{d}y$$

La densité marginale de Y en intégrant $f_{X,Y}$ par rapport à x.

– On définit la variable aléatoire $\mathcal{X} = X|Y$ comme X conditionellement à Y. Sa densité est donnée par :

$$f_{\mathcal{X}}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

– Soit X une variable aléatoire continue dont la densité de probabilité est spécifiée par $f_X(x)$. L'espérance de $\mathcal{G}(X)$, où $\mathcal{G}$ est une fonction continue, est donnée par :

$$\mathbb{E}\left[\mathcal{G}(X)\right] = \int_{-\infty}^{\infty} \mathcal{G}(x) f(x) \mathrm{d}x$$

– Soit (X, Y) un couple de v.a.r. dont la densité de probabilité est spécifiée par $f_{X,Y}(x, y)$. L'espérance de $\mathcal{G}(X, Y)$, où $\mathcal{G}$ est une fonction continue, est donnée par :

$$\mathbb{E}\left[\mathcal{G}(X, Y)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathcal{G}(x, y) f_{X,Y}(x, y) \mathrm{d}x \mathrm{d}y$$

– Notons $\Psi(X) \equiv \mathbb{E}[Y|X]$ l'espérance conditionnelle de Y :

$$\Psi(X) = \int_{-\infty}^{\infty} y \frac{f_{X,Y}(x,y)}{f_X(x)} \mathrm{d}y$$

où le ratio de la densité jointe de (X, Y) et de la densité marginale de X est la densité (conditionnelle) de Y sachant X. L'espérance conditionnelle de Y est une fonction de X, comme le suggère la notation, et donc une variable aléatoire.

THÉORÈME DES ESPÉRANCES ITÉRÉES. L'espérance conditionnelle $\Psi(X) \equiv \mathbb{E}[Y|X]$ satisfait $\mathbb{E}[\Psi(X)] = \mathbb{E}[Y]$.

PREUVE DU THÉORÈME DES ESPÉRANCES ITÉRÉES. Par définition, nous avons :

$$\mathbb{E}\left[\Psi(X)\right] = \int_{\mathbb{R}} \Psi(x) f_X(x) \mathrm{d}x$$

En substituant la définition de l'espérance conditionnelle de Y, il vient :

$$\mathbb{E}\left[\Psi(X)\right] = \int_{\mathbb{R}} \int_{\mathbb{R}} y f_{Y|X}(y|x) f_X(x) \mathrm{d}x \mathrm{d}y$$

ou de façon équivalente :

$$\begin{aligned}\mathbb{E}\left[\Psi(X)\right] &= \int_{\mathbb{R}} y \left(\int_{\mathbb{R}} f_{X,Y}(x,y) \mathrm{d}x \right) \mathrm{d}y \\ &= \int_{\mathbb{R}} y f_Y(y) \mathrm{d}y = \mathbb{E}\left[Y\right]\end{aligned}$$

– Ce théorème nous dit simplement que l'espérance de Y peut s'écrire comme une «moyenne» de l'espérance de $Y|X$ pondérée par la densité de X.
– Ce résultat, utile dans de nombreuses situtations, permet d'évaluer l'espérance d'une variable aléatoire Y, sans explicitement connaître la densité de Y. Il suffit de connaître la densité de $Y|X$ et la densité marginale de X.

THÉORÈME DES ESPÉRANCES ITÉRÉES (GÉNÉRALISATION). L'espérance conditionnelle $\Psi(X) \equiv \mathbb{E}[Y|X]$ satisfait $\mathbb{E}[\Psi(X)\mathcal{G}(X)] = \mathbb{E}[Y\mathcal{G}(X)]$ pour toute fonction $\mathcal{G}$ telle que les deux espérances existent.

Revenons, en utilisant le théorème des espérances itérées, aux hypothèses formulées sur la perturbation du modèle linéaire simple :

– La perturbation est d'espérance nulle :

$$\mathbb{E}\left[\epsilon_i\right] = \mathbb{E}_{\mathbf{x}}\Big[\mathbb{E}\big[\epsilon_i|\mathbf{x}\big]\Big] = \mathbb{E}_{\mathbf{x}}\big[0\big] = 0 \quad \forall i$$

– La perturbation est non corrélée avec la variable explicative :

$$\mathbb{E}\left[\epsilon_i \mathbf{x}\right] = \mathbb{E}_{\mathbf{x}}\Big[\mathbf{x}\mathbb{E}\big[\epsilon_i|\mathbf{x}\big]\Big] = \mathbb{E}_{\mathbf{x}}\Big[\mathbf{x} \times 0\Big] = 0 \quad \forall i$$

– La perturbation est de variance σ^2 :

$$\mathbb{V}\left[\epsilon_i\right] = \mathbb{E}\left[\epsilon_i^2\right] = \mathbb{E}_{\mathbf{x}}\Big[\mathbb{E}\big[\epsilon_i^2|\mathbf{x}\big]\Big] = \mathbb{E}_{\mathbf{x}}\Big[\sigma^2\Big] = \sigma^2 \quad \forall i$$

– Avant de calculer l'estimateur des MCO, notons que dans les hypothèses (i) et (ii) nous conditionnons ϵ_i par rapport à $\mathbf{x}$ c'est-à-dire par rapport à $x_1, x_2, \ldots, x_N$.

– Ainsi, par exemple, $\mathbb{E}[\epsilon_i x_j] = 0$ si $i = j$, $i < j$ ou $j > i$. Cette hypothèse est généralement acceptable lorsque l'indice i n'est pas ordonné (par exemple pour des données en coupe).

– Dans le cas de modèles où l'indice est ordonné, par exemple pour une série temporelle, cette hypothèse est beaucoup trop forte...

– Nous avons défini le DGP. Le modèle estimé est de la forme :

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

– L'esimateur des moindre carrés ordinaires est obtenu en minimisant la somme des carrés des résidus :

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2$$

– Les CNO associées à ce programme sont données par :

$$\sum_{i=1}^{N} \left(y_i - \hat{\alpha} - \hat{\beta} x_i \right) = 0$$

$$\sum_{i=1}^{N} x_i \left(y_i - \hat{\alpha} - \hat{\beta} x_i \right) = 0$$

– L'interprétation de ces deux CNO est directe. Elles nous disent que :

  1. Le résidu est de moyenne nulle.
  2. Le résidu est non corrélé avec la variable explicative.

– En développant la première CNO, on obtient une expression pour l'estimateur de l'ordonnée à l'origine :

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

– En développant la seconde CNO et en substituant l'expression de $\hat{\alpha}$, il vient :

$$\sum_{i=1}^{N} x_i y_i - \left(\bar{y} - \hat{\beta}\bar{x}\right) \sum_{i=1}^{N} -\hat{\beta} \left(N\bar{x}^2 - \sum_{i=1}^{N} x_i^2 \right) = 0$$

– En isolant l'estimateur de β, il vient :

$$\hat{\beta} = \frac{\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - \bar{x}^2}$$

ou de façon équivalente :

$$\hat{\beta} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

– Au dénominateur de l'estimateur de la pente, $\hat{\beta}$, nous retrouvons un estimateur (à un facteur N^{-1} près) de la variance de la variable explicative. Au numérateur nous sommes en présence (toujours à un facteur N^{-1} près) d'un estimateur de la covariance entre la variable expliquée et la variable explicative.

– Ainsi, lorsque la taille de l'échantillon tend vers l'infini :

$$\hat{\beta} \underset{N \to \infty}{\overset{\text{proba}}{\longrightarrow}} \frac{\mathbb{C}ov(X, Y)}{\mathbb{V}[X]}$$

– Notez bien que les estimateurs de α et β dépendent des variables aléatoires $x_1, \ldots, x_N$ et $y_1, \ldots, y_N$. Ces estimateurs sont donc aussi des variables aléatoires réelles.

– $\hat{\beta}$ est-il un estimateur sans biais de b ? Pour répondre à cette question, en comparant l'espérance de $\hat{\beta}$ avec la vraie valeur de la pente b, nous devons utilisez le DGP...

– Le DGP nous dit que :

$$y_i = a + bx_i + \epsilon_i$$

$$\bar{y} = a + b\bar{x} + \bar{\epsilon}$$

– En substituant le DGP dans l'expression de l'esimateur de $\hat{\beta}$, *ie* en remplaçant $y_i - \bar{y}$ dans le numérateur, il vient :

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})((x_i - \bar{x})b + \epsilon_i - \bar{\epsilon})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$= b + \frac{\sum_{i=1}^{N}(x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

– Ainsi, en appliquant l'espérance, nous avons :

$$\mathbb{E}\left[\hat{\beta}\right] - b = \mathbb{E}\left[\frac{\sum_{i=1}^{N}(x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right]$$

L'estimateur est sans biais si et seulement si l'espérance sur le membre de droite est nulle.

– Posons :

$$\Psi_i(\mathbf{x}) = \frac{x_i - \bar{x}}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

– Nous avons alors :

$$\mathbb{E}\left[\hat{\beta}\right] - b = \mathbb{E}\left[\sum_{i=1}^{N}\Psi_i(\mathbf{x})(\epsilon_i - \bar{\epsilon})\right]$$

– Nous pouvons maintenant utiliser le théorème des espérances itérées :

$$\mathbb{E}\left[\hat{\beta}\right] - b = \mathbb{E}_{\mathbf{x}}\left[\sum_{i=1}^{N} \Psi_i(\mathbf{x})\mathbb{E}\left[(\epsilon_i - \bar{\epsilon})|\mathbf{x}\right]\right]$$

– Comme, par hypothèse, l'espérance conditionnelle $\mathbb{E}\left[(\epsilon_i - \bar{\epsilon})|\mathbf{x}\right]$ est nulle, nos avons directement :

$$\mathbb{E}\left[\hat{\beta}\right] = b$$

– $\hat{\beta}$ est un estimateur de la pente b. On établit facilement que $\hat{\alpha}$ est un estimateur sans biais de l'ordonnée à l'origine a.

– Nous avons déjà signalé que $\hat{\beta}$ est une variable aléatoire. La valeur espérée de cette variable aléatoire est la vraie valeur du paramètre que nous cherchons à estimer (b).

– En pratique, pour différents échantillons, l'estimateur des MCO nous donnera différentes estimations distribuées autour de b. Pour évaluer la précision de $\hat{\beta}$ il nous faut calculer la variance de l'estimateur des MCO dans ce modèle. La variance est définie par :

$$\mathbb{V}\left[\hat{\beta}\right] = \mathbb{E}\left[(\hat{\beta} - b)^2\right]$$

car nous avons déjà montré que l'espérance de $\hat{\beta}$ est b.

– Nous avons donc :

$$\begin{aligned}
\mathbb{V}\left[\hat{\beta}\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{N}\Psi_i(\mathbf{x})(\epsilon_i-\bar{\epsilon})\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\Psi_i(\mathbf{x})\Psi_j(\mathbf{x})(\epsilon_i-\bar{\epsilon})(\epsilon_j-\bar{\epsilon})\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\Psi_i(\mathbf{x})\Psi_j(\mathbf{x})\mathbb{E}\left[(\epsilon_i-\bar{\epsilon})(\epsilon_j-\bar{\epsilon})|\mathbf{x}\right]\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\sum_{i=1}^{N}\Psi_i(\mathbf{x})^2\mathbb{E}\left[(\epsilon_i-\bar{\epsilon})^2|\mathbf{x}\right]\right] \\
&= \sigma^2\mathbb{E}\left[\sum_{i=1}^{N}\Psi_i(\mathbf{x})^2\right] = \sigma^2\mathbb{E}\left[\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{\left(\sum_{i=1}^{N}(x_i-\bar{x})^2\right)^2}\right] \\
&= \sigma^2\mathbb{E}\left[\frac{1}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\right]
\end{aligned}$$

– Lorsque la variable explicative est déterministe[a], on retrouve le résultat que vous déjà connaître :

$$\mathbb{V}\left[\hat{\beta}\right] = \sigma^2 \left(\sum_{i=1}^{N} (x_i - \bar{x})^2 \right)^{-1}$$

– La variance de $\hat{\beta}$ est inversement proportionnelle au contenu informationnel de la variable explicative (la variance de x) et proportionnelle à la variance de ϵ (la taille du bruit). Quand une variable explicative est peu variable, on doit s'attendre à obtenir une estimation imprécise du paramètre associé.

[a] Ou en raisonnant conditionnellement à la variable explicative.

– Puisque le terme sous la somme, $(x_i - \bar{x})^2$, est nécessairement positif, notez que lorsque la taille de l'échantillon augmente alors l'inverse de la somme et donc la variance de $\hat{\beta}$ diminue. Asymptotiquement la variance de $\hat{\beta}$ tend vers zéro.

– Ainsi $\hat{\beta}$ tend en probabilité vers b lorsque N tend vers l'infini.

– On peut aussi montrer que $\hat{\alpha}$ tend en probabilité vers a lorsque N tend vers l'infini.