

## II - Estimation et Prévision Ponctuelles

### II.1. Rappels sur l'estimation

- **Modèle**  $(\mathcal{Y}, \mathcal{P}, \Theta)$  ou  $(\mathcal{Y}, \mathcal{X}, \mathcal{P}, \Theta)$   
Paramètre  $\theta$  scalaire ( $\Theta = \mathbb{R}$ ) ou vectoriel ( $\Theta = \mathbb{R}^K$ )
- **Estimateur** :  $\theta^* = f(y_1, \dots, y_N; x_1, \dots, x_N)$
- **Biais** =  $\theta - E_\theta(\theta^*)$

$\theta^*$  est dit sans biais si quel que soit  $\theta$  :

$$E_\theta(\theta^*) = \theta$$

7

L'estimateur  $\theta^*$  est fonction de variables aléatoires, il est lui aussi aléatoire.

$E_\theta$  désigne l'espérance mathématique selon une loi associée à la valeur  $\theta$  du paramètre.

Le biais n'est défini que si  $\theta^*$  admet une espérance mathématique finie.

Exemple :

Soient  $y_1, y_2, \dots, y_N$  i.i.d. loi de Poisson de paramètre  $\lambda$  (loi des événements rares et indépendants : par exemple,  $y_i$  = nombre de propositions d'embauche reçues durant le  $i$ -ème mois).

$Y = \mathbb{N}^N$ ,  $\mathcal{P}$  = l'ensemble des lois de Poisson,  $\Theta = \{\text{réels} > 0\}$

Soit l'estimateur  $\lambda^* = \sum y_n / N$ , moyenne empirique des  $y_n$ .

Pour tout  $\lambda$ ,  $E_\lambda[\lambda^*] = \sum E_\lambda[y_i] / n = n \lambda / n = \lambda$ .

$\Rightarrow$  l'estimateur  $\lambda^*$  est sans biais

$\text{Var}[\lambda^*] = \sum V_\lambda[y_i] / n^2 = n \lambda / n^2 = \lambda / n$ .

Le risque quadratique moyen est  $\text{RQM} = 0 + \lambda / n = \lambda / n$

## II.1 Rappels sur l'estimation (suite...)

### • Risque Quadratique Moyen

$$\text{RQM} = E_{\theta} \|\theta - \theta^*\|^2$$

$$\text{RQM} = \|\text{biais}\|^2 + \Sigma V(\theta_i^*)$$

Si  $\theta$  est scalaire, le RQM est minimum si  $\theta^*$  est sans biais et de variance minimum (cela n'est plus vrai pour  $\theta$  vectoriel).

8

$\theta$  est non aléatoire,  $\theta^*$  est aléatoire.  $E_{\theta}$  et  $V_{\theta}$  désignent l'espérance et la variance calculées pour une loi associée à la valeur  $\theta$ .

Pour un paramètre scalaire :

$$\text{RQM} = E_{\theta} [\theta - \theta^*]^2 = E_{\theta} [\theta - E_{\theta}(\theta^*) + E_{\theta}(\theta^*) - \theta^*]^2$$

$$\text{RQM} = E_{\theta} \{ [\theta - E_{\theta}(\theta^*)]^2 + [E_{\theta}(\theta^*) - \theta^*]^2 + 2 [\theta - E_{\theta}(\theta^*)] [E_{\theta}(\theta^*) - \theta^*] \}$$

$$\bullet E_{\theta} \{ [\theta - E_{\theta}(\theta^*)]^2 \} = [\theta - E_{\theta}(\theta^*)]^2 \text{ (car tout est certain dans l'expression)}$$

$$E_{\theta} \{ [\theta - E_{\theta}(\theta^*)]^2 \} = \text{biais}^2$$

$$\bullet E_{\theta} \{ [E_{\theta}(\theta^*) - \theta^*]^2 \} = V_{\theta}[\theta^*] \text{ (définition de la variance)}$$

$$\bullet E_{\theta} \{ [\theta - E_{\theta}(\theta^*)] [E_{\theta}(\theta^*) - \theta^*] \} = [\theta - E_{\theta}(\theta^*)] E_{\theta} \{ E_{\theta}(\theta^*) - \theta^* \}$$

$$E_{\theta} \{ [\theta - E_{\theta}(\theta^*)] [E_{\theta}(\theta^*) - \theta^*] \} = [\theta - E_{\theta}(\theta^*)] [E_{\theta}(\theta^*) - E_{\theta}(\theta^*)] = 0$$

$$\text{RQM} = \text{biais}^2 + V_{\theta}[\theta^*]$$

Si  $\theta$  est vectoriel, de dimension  $K$ , le risque quadratique moyen est défini par

$$\text{RQM} = E_{\theta} (\theta - \theta^*)'(\theta - \theta^*) = E_{\theta} \Sigma_k (\theta_k - \theta_k^*)^2 = \Sigma_k E_{\theta} (\theta_k - \theta_k^*)^2$$

$$\text{RQM} = \Sigma_k [\text{biais}_k^2 + V_{\theta}(\theta_k^*)]$$

$$\text{RQM} = \|\text{biais}\|^2 + \Sigma_k V_{\theta}[\theta_k^*]$$

En anglais, RQM se traduit par

MSE = Mean Square Error

## II.1 Rappels sur l'estimation (suite...)

- **Convergence**

(en probabilité, presque sûre ou en moyenne quadratique)

$\theta^*$  est convergent si il « tend » vers la vraie valeur  $\theta$  du paramètre lorsque  $N$  tend vers l'infini (limite « en probabilité » ou « presque sûre » ou « en moyenne quadratique »)

- **Condition suffisante de convergence**

Pour que  $\theta^*$  soit convergent, il suffit que :

$$E_{\theta}(\theta^*) \rightarrow \theta \text{ et } V_{\theta}(\theta^*) \rightarrow 0.$$

L'estimateur est une fonction des observations : à chaque taille  $N$  d'échantillon, il correspond un estimateur  $\theta_N^*$ . Nous avons donc une suite de variables aléatoires. La convergence d'une suite de variables aléatoires est définie de différentes manières. Nous considérons ici la convergence en probabilité.

Convergence en probabilité :

$P[|\theta_N^* - \theta| < \epsilon]$  tend vers 1 quel que soit  $\epsilon > 0$ .

## II. 2 Méthodes d'estimation

### II.2.a. Maximum de vraisemblance dans un modèle paramétrique

**Vraisemblance du paramètre** = loi de probabilité des observations, considérée comme fonction du paramètre :

$$f(y_1, y_2, \dots, y_N; \theta) = \ell(\theta)$$

**Estimateur du maximum de vraisemblance** = solution de  $Max_{\theta} [\ell(\theta)]$

10

Exemple : n-échantillon de loi de Bernoulli (indicateur d'un événement, vaut 1 si l'événement est réalisé et 0 sinon)

$$Y \approx B(1;p) \quad P(Y = y) = p^y (1-p)^{1-y} \quad \text{pour } y = 1 \text{ ou } 0$$

$$\begin{aligned} \ell(y_1, y_2, \dots, y_N; p) &= \prod_{i=1}^N [p^{y_i} (1-p)^{1-y_i}] = p^{\sum_{i=1}^N y_i} (1-p)^{N - \sum_{i=1}^N y_i} \\ \ln \ell(y_1, y_2, \dots, y_N; p) &= \sum_{n=1}^N y_n \ln(p) + \left( N - \sum_{n=1}^N y_n \right) \ln(1-p) \\ L(y_1, y_2, \dots, y_N; p) &= N \left[ \bar{y} \ln(p) + (1-\bar{y}) \ln(1-p) \right] \end{aligned}$$

Le maximum de vraisemblance a la même solution que le maximum de la logvraisemblance. Les dérivées de la logvraisemblance sont :

$$\begin{aligned} \frac{\partial L}{\partial p} &= n \left[ \frac{\bar{y}}{p} - \frac{1-\bar{y}}{1-p} \right] = \frac{n[\bar{y}-p]}{p(1-p)} \\ \frac{\partial^2 L}{\partial p^2} &= -n \left[ \frac{\bar{y}}{p^2} + \frac{1-\bar{y}}{(1-p)^2} \right] < 0 \end{aligned}$$

La dérivée par rapport à p est nulle pour

$$\hat{p} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \text{fréquence observée}$$

La dérivée seconde est négative :

la solution correspond à un maximum

## II.2.b. Méthode des moments

Pour la loi  $L(y;\theta)$ , moments :  $m_1 = E_0[y]$ ,  $m_2 = E_0[y^2]$

**Moments fonction de  $\theta$**

$$m_1 = f_1(\theta_1, \theta_2)$$

$$m_2 = f_2(\theta_1, \theta_2)$$

**$\theta$  fonction des moments**

$$\theta_1 = g_1(m_1, m_2)$$

$$\theta_2 = g_2(m_1, m_2)$$

Observation d'un N-échantillon  $(y_1, \dots, y_N)$

**Moments empiriques :**

$$\hat{m}_1 = \frac{1}{N} [y_1 + \dots + y_N]$$

$$\hat{m}_2 = \frac{1}{N} [y_1^2 + \dots + y_N^2]$$

**Estimateur de  $\theta$**

$$\hat{\theta}_1 = g_1(\hat{m}_1, \hat{m}_2)$$

$$\hat{\theta}_2 = g_2(\hat{m}_1, \hat{m}_2)$$

11

Exemple : Un assureur étudie la distribution du coût d'un certain sinistre pour un certain type d'assuré. Un modèle classique attribue à ce coût  $Y$  une loi Gamma, de densité :

$$f(y; a, r) = \begin{cases} \frac{1}{a^r \Gamma(r)} y^{r-1} e^{-y/a} & \text{si } y > 0 \\ 0 & \text{sinon} \end{cases}$$

où  $r$  et  $a$  sont des paramètres strictement positifs.

On peut montrer (car  $2Y/a$  suit un chi-deux à  $2r$  degrés de liberté) que

$$m_1 = E(Y) = r a$$

$$V = V(Y) = r a^2$$

On en déduit : 
$$a = \frac{V(Y)}{E(Y)} \quad \text{et} \quad r = \frac{[E(Y)]^2}{V(Y)}$$

Les moments empiriques sont, pour un échantillon de taille  $n$

$$\hat{m}_1 = \frac{1}{n} \sum_{i=1}^n y_i \quad \text{et} \quad \hat{V} = \frac{1}{n} \sum_{i=1}^n y_i^2 - \hat{m}_1^2$$

L'estimateur par la méthode des moments est donc :

$$\hat{a} = \frac{\hat{V}}{\hat{m}_1} \quad \text{et} \quad \hat{r} = \frac{[\hat{m}_1]^2}{\hat{V}}$$

## Exemple 1 : coût de sinistres

$$f(y; a, r) = \begin{cases} \frac{1}{a^r \Gamma(r)} y^{r-1} e^{-y/a} & \text{si } y > 0 \\ 0 & \text{sinon} \end{cases}$$

loi gamma, paramètres  $a$  et  $r$  positifs

$$\begin{cases} E(Y) =: m = ra \\ V(Y) =: V = ra^2 \end{cases} \Leftrightarrow \begin{cases} a = \frac{V}{m} \\ r = \frac{m^2}{V} \end{cases}$$

Moments empiriques  $\Rightarrow$  estimateur de  $a$  et  $r$  :

$$\begin{aligned} \hat{m} &= \frac{1}{N} \sum_{i=1}^N y_i \\ \hat{V} &= \frac{1}{N} \sum_{i=1}^N (y_i - \hat{m})^2 \end{aligned} \quad \begin{cases} \hat{a} = \frac{\hat{V}}{\hat{m}} \\ \hat{r} = \frac{\hat{m}^2}{\hat{V}} \end{cases}$$

12

Remarques :

- Il est possible d'utiliser les moments centrés et les moments non centrés. Ici, la variance est plus pratique à utiliser que le moment non centré d'ordre 2. Le principe de la méthode est inchangé.
- Les moments centrés et non centrés sont liés par des relations. En ce qui concerne la variance :

Variance (théorique) :

$$V(Y) = E(Y^2) - [E(Y)]^2$$

Variance empirique :

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

## Exemple 2 : durées d'attente

$$f(y; \theta) = \begin{cases} \frac{1}{\theta} & \text{si } 0 < y < \theta \\ 0 & \text{sinon} \end{cases}$$

loi uniforme sur  $[0, \theta]$ , paramètre  $\theta$  positif

$$E(Y) =: m = \frac{\theta}{2} \quad \Leftrightarrow \quad \theta = 2m$$

Moyenne empirique  $\Rightarrow$  estimateur de  $\theta$  :

$$\hat{m} = \frac{1}{N} \sum_{i=1}^N y_i \quad \hat{\theta} = 2\hat{m}$$

13

Noter que l'estimateur des moments est ici différent de l'estimateur du maximum de vraisemblance : la vraisemblance de  $\theta$  est ici :

$$\ell(\theta) = \frac{1}{\theta^N} \text{ si } \{\theta > x_1, \dots, \theta > x_N\} \text{ et } 0 \text{ sinon}$$

$$\ell(\theta) = \frac{1}{\theta^N} \text{ si } \theta > \text{Max}\{x_1, \dots, x_N\}$$

Cette vraisemblance est une fonction décroissante de  $\theta$ . Son maximum est atteint pour le plus petit  $\theta$  possible, c'est à dire  $\text{Max}(x_1, \dots, x_N)$  (le maximum de vraisemblance est atteint sur le bord du support de la vraisemblance, la dérivée n'y est pas nulle). L'estimateur du maximum de vraisemblance est donc :

$$\hat{\theta}_{MV} = \text{Max}\{x_1, \dots, x_N\}$$

## II.2.c. Ajustement par les Moindres Carrés Ordinaires

Dans la famille des fonctions :  $f(x ; \theta)$ ,  $\theta \in \Theta$ , la fonction "la plus proche" des valeurs observées  $(y_1, x_1), (y_2, x_2), \dots, (y_N, x_N)$ , est celle correspondant à la solution  $\hat{\theta}$  de :

$$\text{Min}_{\theta} \left\{ \sum_{i=1}^N [y_i - f(x_i; \theta)]^2 \right\}$$

$\hat{y}_i = f(x_i; \hat{\theta})$  est la *valeur ajustée* par les MCO

14

Les fonctions utilisées pour de tels ajustements sont :

fonctions exponentielles , droites (ou fonctions affines), etc...

Nous traiterons en particulier l'ajustement MCO à une droite.

La fonction solution est dite ajustée par les MCO.

La valeur optimale  $\hat{\theta}$  de  $\theta$  est l'estimateur MCO de  $\theta$ .

Les erreurs  $\hat{u}_i = y_i - f(x_i; \hat{\theta})$  sont les résidus MCO

et la valeur minimum atteinte est la somme des carrés des résidus, notée SCR



## (i) Cas du modèle linéaire sans constante

le modèle

$$y_i = a x_i + u_i, \quad i = 1, \dots, N$$
$$x_i \text{ certaines, } E(u_i) = 0$$

la famille de fonctions :  $\{y = ax, a \text{ réel}\}$

L'estimateur MCO de  $a$  est solution de :

$$\text{Min}_a \left[ (y_1 - a x_1)^2 + \dots + (y_N - a x_N)^2 \right]$$

$$\text{c'est : } \hat{a} = \frac{x_1 y_1 + \dots + x_N y_N}{x_1^2 + \dots + x_N^2} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

15

La condition nécessaire d'ordre 1 s'écrit en dérivant la fonction objectif par rapport à  $a$  :

$$\frac{\partial}{\partial a} \left[ (y_1 - a x_1)^2 + \dots + (y_n - a x_n)^2 \right] = -2x_1(y_1 - a x_1) - \dots - 2x_n(y_n - a x_n) = 0$$

d'où on tire :

$$\hat{a} = \frac{x_1 y_1 + \dots + x_n y_n}{x_1^2 + \dots + x_n^2} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

Condition suffisante d'ordre 2 :

$$\frac{\partial^2}{\partial a^2} \left[ (y_1 - a x_1)^2 + \dots + (y_n - a x_n)^2 \right] = 2x_1^2 + \dots + 2x_n^2 > 0$$

La solution trouvée correspond bien à un minimum.

Propriétés de cet estimateur : on peut écrire  $\hat{a}$  sous la forme suivante

$$\hat{a} = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{\sum_i x_i (a x_i + u_i)}{\sum_i x_i^2} = a + \frac{\sum_i x_i u_i}{\sum_i x_i^2}$$

D'où l'on tire facilement que

$$E(\hat{a}) = a$$

L'estimateur MCO dans un modèle linéaire est donc sans biais.

Pour calculer sa variance, il faut faire des hypothèses supplémentaires sur les variances et covariances des perturbations.

## (ii) Cas du modèle linéaire avec constante

le modèle :

$$y_i = a x_i + b + u_i, \quad i = 1, \dots, N$$

$$x_i \text{ certaines, non constantes, } E(u_i) = 0$$

la famille de fonctions :  $\{y = ax + b ; a, b \text{ réels}\}$

L'estimateur MCO de  $(a, b)$  est solution de :

$$\text{Min}_a \left[ (y_1 - a x_1 - b)^2 + \dots + (y_N - a x_N - b)^2 \right]$$

$$\text{c'est : } \hat{a} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_n (x_i - \bar{x})^2} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

$$\text{où } \bar{x} = \frac{1}{N} \sum_i x_i \text{ et } \bar{y} = \frac{1}{N} \sum_i y_i$$

16

La condition nécessaire d'ordre 1 s'écrit en dérivant la fonction objectif par rapport à  $a$  et  $b$ :

$$\frac{\partial}{\partial a} \left[ (y_1 - a x_1 - b)^2 + \dots + (y_N - a x_N - b)^2 \right] = -2x_1(y_1 - a x_1 - b) - \dots - 2x_N(y_N - a x_N - b) = 0$$

$$\frac{\partial}{\partial b} \left[ (y_1 - a x_1 - b)^2 + \dots + (y_N - a x_N - b)^2 \right] = -2(y_1 - a x_1 - b) - \dots - 2(y_N - a x_N - b) = 0$$

d'où on tire :

$$\begin{cases} a \sum_i x_i^2 + b \sum_i x_i = \sum_i x_i y_i \\ a \sum_i x_i + bN = \sum_i y_i \\ a \left[ \frac{1}{N} \sum_i x_i^2 - (\bar{x})^2 \right] = \frac{1}{N} \sum_i x_i y_i - \bar{x}\bar{y} \\ a\bar{x} + b = \bar{y} \end{cases}$$

Facultatif : la Condition suffisante d'ordre 2 fait intervenir les dérivées secondes

$$\frac{\partial^2}{\partial a^2} \left[ (y_1 - a x_1 - b)^2 + \dots + (y_N - a x_N - b)^2 \right] = 2 \sum_i x_i^2$$

$$\frac{\partial^2}{\partial a \partial b} \left[ (y_1 - a x_1 - b)^2 + \dots + (y_N - a x_N - b)^2 \right] = 2 \sum_i x_i$$

$$\frac{\partial^2}{\partial b^2} \left[ (y_1 - a x_1 - b)^2 + \dots + (y_N - a x_N - b)^2 \right] = 2N$$

La matrice des dérivées secondes est de format  $2 \times 2$ . Son déterminant est  $> 0$ , ainsi que sa trace (somme des termes diagonaux). Elle est donc définie positive. La solution trouvée correspond bien à un minimum.

## Une étude de Monte Carlo

- Sont données 22 valeurs d'une variable  $X$  :

$$x_1, x_2, \dots, x_{22}$$

- Nous tirons un échantillon de taille 22 d'une loi Normale  $N(0 ; 1)$ , et nous calculons les valeurs  $y_i$  :

$$y_i = ax_i + b + u_i .$$

Nous obtenons un échantillon de taille 22 de  $Y$ , suivant un modèle linéaire avec constante.

- Répétons l'expérience : le paramètre  $(a, b)$  est toujours le même, ainsi que les  $x_i$ , mais les  $y_i$  varient d'un tirage à l'autre, ainsi que les estimations de  $a$  et de  $b$

17

On vérifie expérimentalement le caractère certain des  $x$  et de  $(a, b)$  et le caractère aléatoire des  $y$  et des estimations faites.

On vérifie que les valeurs  $y$  et leurs ajustements sont de même moyenne empirique

Cas du modèle linéaire avec constante (suite)

**Proposition.** Dans le cadre du modèle linéaire défini ci-dessus, l'estimateur MCO est sans biais

$$E(\hat{A}) = a, \quad E(\hat{B}) = b$$

18

Le calcul des estimateurs et de leurs propriétés est plus facile quand on utilise les notations et calculs matriciels.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \text{ dans } \mathbb{R}^N$$

modèle LS :  $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$

$$V(\mathbf{y}) = \sigma^2 \mathbf{I}_N$$

estimateur MCO solution des équations normales :

$$\mathbf{X}'\mathbf{X} \boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$$

$$\boldsymbol{\beta}^* = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

$$E(\boldsymbol{\beta}^*) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} : \text{estimateur sans biais}$$

$$V(\boldsymbol{\beta}^*) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'V(\mathbf{y}) \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}, \text{ d'où les formules.}$$

$$X'X = \begin{pmatrix} \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & N \end{pmatrix}, \quad (X'X)^{-1} = \frac{1}{N \sum_{i=1}^N (x_i - \bar{x})^2} \begin{pmatrix} N & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{pmatrix}$$

Formules à rappeler :  $\sum_{i=1}^N x_i y_i - N\bar{x}\bar{y} = \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$

La condition suffisante d'ordre 2 qui porte sur la matrice des dérivées secondes se démontre aussi plus simplement et le raisonnement est généralisable au cas de plus de deux régresseurs

$$H(a, b) = 2 \begin{pmatrix} \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & N \end{pmatrix} = 2X'X ; \text{ donc pour tout } b : b'X'Xb = (Xb)'Xb = \|Xb\|^2 \geq 0$$

## Calcul des Variances-covariances

Modèle linéaire standard = modèle linéaire plus hypothèse sur les variances et covariances des  $u_i$  :  $V(u_i) = \sigma^2$ , et  $\text{cov}(u_i, u_j) = 0$  si  $i$  est différent de  $j$ .

- Dans le modèle sans constante :

$$\hat{A} = \frac{\sum_i x_i Y_i}{\sum_i x_i^2} : V(\hat{A}) = \frac{\sigma^2}{\sum_i x_i^2}$$

- Dans le modèle avec constante :  $V(\hat{\beta}) = \sigma^2 (X'X)^{-1}$

$$V(\hat{A}) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}, V(\hat{B}) = \frac{\sigma^2 \sum_i x_i^2}{\sum_i (x_i - \bar{x})^2}$$

$$\text{cov}(\hat{A}, \hat{B}) = -\frac{\sigma^2 \bar{x}}{\sum_i (x_i - \bar{x})^2}$$

19

Dans le cas du modèle sans constante,  $\mathbf{X} = (\mathbf{x})$  a une seule colonne, et  $\mathbf{X}'\mathbf{X}$  est le scalaire  $\sum x_i^2$ .

Dans le cas du modèle avec constante,  $\mathbf{X} = (\mathbf{x} \ \mathbf{e})$  et  $\mathbf{X}'\mathbf{X}$  est une matrice 2\*2. La variance de  $\hat{a}$  est  $\sigma^2$  fois l'élément 1ère ligne, 1ère colonne de  $(\mathbf{X}'\mathbf{X})^{-1}$ , etc : les formules s'en déduisent immédiatement.

## Théorème de GAUSS-MARKOV

Sous les hypothèses du **modèle linéaire standard**, l'estimateur MCO est le meilleur estimateur de  $(a, b)$  parmi les estimateurs linéaires et sans biais (BLUE)

- **BLUE** = Best Linear Unbiased Estimator
- "meilleur" veut dire que pour toute combinaison  $\theta = a x_0 + b$ , l'estimateur de variance minimum est :

$$\hat{\theta} = \hat{a} x_0 + \hat{b}$$

20

Pour  $\theta$  scalaire,  $\theta_1$  est meilleur que  $\theta_2$  si  $V(\theta_1) \leq V(\theta_2)$

Pour  $\theta \in \mathbb{R}^k$ ,  $\theta_1$  est meilleur que  $\theta_2$  si la propriété est vraie pour toute combinaison linéaire de leurs composantes :

Pour tout  $c$  de  $\mathbb{R}^k$ :  $V(c'\theta_1) \leq V(c'\theta_2)$

$c$  'est-à-dire  $c'V(\theta_1)c \leq c'V(\theta_2)c$

ou encore  $c'[V(\theta_1) - V(\theta_2)]c \leq 0$

(en math, on dit que la matrice  $V(\theta_1) - V(\theta_2)$  est semi-définie négative)

Par exemple, pour  $K = 2$  :

quel que soit  $x_t$ , la meilleure estimation de  $E(y_t) = ax_t + b$

est  $\hat{E}(y_t) = \hat{a}x_t + \hat{b}$  où  $\hat{a}$  et  $\hat{b}$  sont les estimations MCO de  $a$  et  $b$ .

## II.3. Prévision

La prévision de la variable aléatoire  $y$ , notée  $y^*$ , est calculée en fonction de l'information disponible.

- On suppose ici que  $y$  est sans corrélation avec les observations déjà faites.
- On note  $m$  l'espérance mathématique de  $y$

21

Exemple : modèle linéaire  $y_t = a x_t + u_t$   $x_t$  certaines,  $E(u_t) = 0$ ,  $a$  réel pour  $x$  donnée, on suppose que  $y = a x + u$ , où  $y$  est sans corrélation avec les valeurs déjà observées.

La meilleure prévision de  $y$  serait  $a x$  si on connaissait  $a$ .

Puisque  $a$  est inconnu, on prend comme prévision de  $y$  la variable

$$y^* = \hat{a} x$$

Problème : est-ce bien la meilleure prévision?

Réponse : oui, si les perturbations vérifient les conditions standard d'ordre 2 = homoscedasticité et non-corrélation les unes avec les autres (Th. de Gauss-Markov)

Dans ce cas,  $E(\hat{a}) = a$  et  $V(\hat{a}) = \sigma^2 / \sum x_t^2$

$$\text{erreur} = y - y^* = a x + u - \hat{a} x = u + (a - \hat{a}) x$$

$$E(\text{erreur}) = E(u) + E(a - \hat{a}) x = 0$$

$$V(\text{erreur}) = V(u) + V(\hat{a}) x^2 + 2 \text{cov}(u, \hat{a}) = \sigma^2 + \sigma^2 x^2 / \sum x_t^2 + 0$$

### II.3. Prévision

- Erreur quadratique moyenne (MSE)

$$\text{MSE} = E(y - y^*)^2 = V(y) + E(m - y^*)^2$$

- Prévision théorique :  $m$  connue,  $y^* = m$
- Prévision empirique :  $m$  inconnue,  $y^* = m^*$  est l'estimateur de  $m$  qui minimise le Risque Quadratique Moyen  $E(m - y^*)^2$

22

Dans le cas du modèle linéaire précédent, nous avons :

$$E(\hat{a}) = a \quad \text{et} \quad V(\hat{a}) = \sigma^2 / \sum x_t^2$$

$$\text{erreur} = y - y^* = a x + u - \hat{a} x = u + (a - \hat{a}) x$$

$E(\text{erreur}) = E(u) + E(a - \hat{a}) x = 0$  : on dit que la prévision est sans biais.

$$V(\text{erreur}) = V(u) + V(\hat{a}) x^2 + 2 \text{cov}(u, \hat{a}) = V(u) + V(\hat{a}) x^2$$

L'erreur quadratique moyenne est

$$\text{MSE} = \sigma^2 + \sigma^2 x^2 / \sum x_t^2$$

Une partie est due à l'incertitude sur  $y$  qui est aléatoire, la deuxième partie est due à l'incertitude sur la loi de  $y$ , c'est-à-dire sur le paramètre  $a$ .



## Exemple 1

- Nous observons, sur un échantillon de 100 étudiants à la recherche d'un emploi au sortir de leur DEA, une durée moyenne d'inactivité de 15,65 semaines.
- On suppose que la durée d'inactivité suit une loi exponentielle de paramètre  $\theta$
- Sortant du même DEA, vous cherchez un emploi : quelle prévision faites-vous du temps qui vous sera nécessaire pour que votre recherche aboutisse ?

23

Nous disposons d'un échantillon de taille 100 d'une variable exponentielle de densité  $f(x) = 1/a \exp(-x/a)$  pour  $x > 0$ . Le meilleur estimateur de  $a$  est la moyenne empirique de l'échantillon.

Si vous supposez que la durée de votre recherche a la même loi que celle des étudiants déjà observés et qu'elle est indépendante des observations passées, alors la meilleure prévision de la durée  $y$  de votre recherche est cette estimation de  $a$ .

Prévision optimale  $\hat{a} = \Sigma x_i / 100$

Erreur de prévision :  $y - \hat{a} = y - a + a - \hat{a}$

$MSE = E(y - a)^2 + E(a - \hat{a})^2 = V(y) + V(\hat{a})$

$MSE = a^2 + a^2/100$

application numérique : vous prévoyez  $y^* = 15,65$  semaines de recherche

## Exemple 2

- Nous sommes dans le cadre d'un modèle linéaire standard où

$$y_i = a x_i + b + u_i, \quad i = 1, \dots, N$$

$x_i$  certaines, non constantes,

$$E(u_i) = 0$$

$$V(u_i) = \sigma^2, \quad \text{cov}(u_i, u_j) = 0 \text{ pour } i \neq j$$

- Quelle est la prévision de  $y_o$  pour un individu indépendant des  $N$  observés et pour lequel nous connaissons  $x_o$  ?
- Quelle sera l'erreur quadratique moyenne correspondante ?

24

Nous reprenons ici les termes du modèle utilisé pour la simulation faite.

La meilleure prévision serait  $E(y_o) = ax_o + b$  si nous connaissions  $a$  et  $b$ . L'erreur correspondante serait de variance  $\sigma^2$ .

Pour utiliser les observations précédentes, nous devons supposer que le paramètre est toujours le même pour le nouvel individu (notamment  $x_o$  ne doit pas être trop éloigné des valeurs déjà observées).

Le modèle étant linéaire standard, le théorème de Gauss-Markov nous dit que la meilleure estimation (linéaire) de  $a$  et  $b$  sont données par l'estimation MCO.

## Exemple 2 (suite)

- Nous devons également supposer que le paramètre n'a pas changé pour le nouvel individu

$$y_o = a x_o + b + u_o ,$$

$$E(u_o) = 0$$

$$V(u_o) = \sigma^2, \text{ cov}(u_o, u_j) = 0 \text{ pour } j = 1, \dots, N$$

- La prévision optimale est obtenue à partir des estimateurs MCO :  $\hat{y}_o = \hat{a} x_o + \hat{b}$

- Cette prévision est sans biais :

$$y_o - \hat{y}_o = a x_o + b + u_o - \hat{a} x_o - \hat{b} = u_o + (a - \hat{a})x_o + b - \hat{b}$$

$$E(y_o - \hat{y}_o) = 0$$

## Exemple 2 (suite)

- L'erreur quadratique moyenne est :

$$MSE = E(u_o + (a - \hat{a})x_o + b - \hat{b})^2 = E(u_o + (a - \hat{a})x_o + b - \hat{b})^2$$

$$MSE = E(u_o)^2 + E((a - \hat{a})x_o + b - \hat{b})^2 + 2 \text{cov}(u_o, (a - \hat{a})x_o + b - \hat{b})$$

$$MSE = \sigma^2 + x_o^2 V(\hat{a}) + V(\hat{b}) + 2 x_o \text{cov}(\hat{a}, \hat{b}) + 0$$

Application numérique :

$$\hat{y}_i = 0,909 x_i + 2,173 ; x_o = 12$$

$$\sigma^2 = 1 \text{ (supposée connue).}$$

$$(X'X)^{-1} = \begin{bmatrix} 0,00271137 & -0,02327467 \\ -0,02327467 & 0,24524639 \end{bmatrix}$$

$$\hat{y}_o = 13,08$$

$$MSE = 1,077$$

26

La MSE est en fait la variance de l'erreur de prévision, puisque cette erreur est d'espérance nulle.

Faire bien attention de repérer dans ces expressions ce qui est aléatoire et ce qui ne l'est pas : les variables aléatoires sont la perturbation  $u_o$  et les estimateurs.

Si on fait un nouveau tirage des  $y_i$ , MSE n'est pas changée (les paramètres ne changent pas et les  $x_i$  sont fixées), mais la prévision change avec les estimations de  $a$  et  $b$ .

Pour le calcul de la MSE, nous avons

$$V(\hat{a}) = \sigma^2 0,00271137$$

$$\text{cov}(\hat{a}, \hat{b}) = -\sigma^2 0,02327467$$

$$V(\hat{b}) = \sigma^2 0,24524639$$

Nous avons supposé que la variance de  $y_i$  est connue, et égale à 1.

Nous verrons plus tard comment faire pour estimer cette variance.