Document disponible à http://www.univ-montp3.fr/miap/ens/AES/L1/optionmath.

Université Paul Valéry
Arts - Lettres - Langues
Sciences humaines & sociales

AES option mathématique
Année 2004–2005

# Notes de cours : ajustement linéaire

## 1 Cadre : mesure conjointe de deux caractères

On se place dans le cas où, sur *une seule* population, on étudie *deux caractères* quantitatifs dans le but d'exhiber un lien entre ces deux caractères.

Soit $X$ le premier caractère et $Y$ le second. On note $\{m_1, m_2, \dots, m_k\}$ les modalités de $X$ et $\{m'_1, m'_2, \dots, m'_\ell\}$ les modalités de $Y$. Pour un couple de modalités $(m_i, m'_j)$, on note $n_{ij}$ l'effectif des individus associés à la modalité $m_i$ pour $X$ *et* $m'_j$ pour $Y$. La somme de toutes les valeurs $n_{ij}$ lorsque $(i, j)$ parcourt $\{1, \dots, k\} \times \{1, \dots, \ell\}$ est donc l'effectif total qu'on note $n$ :

$$\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij} = n. \tag{1}$$

La fréquence associée au couple de modalités $(m_i, m_j)$ est la proportion, parmi toute la population, des individus associés à la modalité $m_i$ pour $X$ et $m'_j$ pour $Y$. On note $f_{ij}$ cette fréquence et on a

$$f_{ij} = \frac{n_{ij}}{n}.$$

On déduit, en divisant par $n$ l'équation (1) que la somme de toutes les valeurs $n_{ij}$ lorsque $(i, j)$ parcourt $\{1, \dots, k\} \times \{1, \dots, \ell\}$ vaut 1 :

$$\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} f_{ij} = 1.$$

## 2 Tableau de contingence

Les résultats sont généralement représentés dans un *tableau de contingence* en effectifs (ou en fréquences) : à chaque modalité de $X$, on associe une ligne, à chaque modalité de $Y$, on associe une colonne puis à l'intersection de la ligne associée à $m_i$ et de la colonne associée à $m'_j$, on place l'effectif $n_{ij}$ (ou la fréquence $f_{ij}$) – voir les tableaux 1 et 2 page suivante.

| | $m'_1$ | … | $m'_j$ | … | $m'_\ell$ |
|---|---|---|---|---|---|
| $m_1$ | $n_{11}$ | | $n_{1j}$ | | $n_{1\ell}$ |
| ⋮ | | | | | |
| $m_i$ | $n_{i1}$ | | $n_{ij}$ | | $n_{i\ell}$ |
| ⋮ | | | | | |
| $m_k$ | $n_{k1}$ | | $n_{kj}$ | | $n_{k\ell}$ |

TAB. 1 – Tableau de contingence en effectifs

| | $m'_1$ | … | $m'_j$ | … | $m'_\ell$ |
|---|---|---|---|---|---|
| $m_1$ | $f_{11}$ | | $f_{1j}$ | | $f_{1\ell}$ |
| ⋮ | | | | | |
| $m_i$ | $f_{i1}$ | | $f_{ij}$ | | $f_{i\ell}$ |
| ⋮ | | | | | |
| $m_k$ | $f_{k1}$ | | $f_{kj}$ | | $f_{k\ell}$ |

TAB. 2 – Tableau de contingence en fréquences

# 3 Moyennes et variances

À partir des observations conjointes de $X$, on peut calculer la moyenne $\mathrm{Moy}(X)$ de $X$ et la moyenne $\mathrm{Moy}(Y)$ de $Y$.

Étant donné une modalité $m_i$ de $X$, un individu associé à cette modalité est associé à une (et une seule) des modalités de $Y$. Ainsi, l'effectif, $n_{i\cdot}$, des individus associés à la modalité $m_i$ de $X$ est la somme des effectifs associés aux couples de modalités $(m_i, m'_j)$ lorsque $j$ parcourt $\{1,\dots,\ell\}$ :

$$n_{i\cdot} = \sum_{j=1}^{\ell} n_{ij} = n_{i1} + \cdots + n_{i\ell}.$$

On rappelle que la moyenne de $X$ est la somme des modalités de $X$ multipliée par les effectifs correspondant divisée par l'effectif total, ainsi :

$$\mathrm{Moy}(X) = \frac{1}{n}\sum_{i=1}^{k} n_{i\cdot} m_i = \frac{n_{1\cdot} m_1 + \cdots + n_{k\cdot} m_k}{n}. \tag{2}$$

Étant donné une modalité $m'_j$ de $Y$, un individu associé à cette modalité est associé à une (et une seule) des modalités de $X$. Ainsi, l'effectif, $n_{\cdot j}$, des individus associés à la modalité $m'_j$ de $Y$ est la somme des effectifs associés aux couples de modalités $(m_i, m'_j)$ lorsque $i$ parcourt $\{1,\dots,k\}$ :

$$n_{\cdot j} = \sum_{i=1}^{k} n_{ij} = n_{1j} + \cdots + n_{kj}$$

puis

$$\mathrm{Moy}(Y) = \frac{1}{n}\sum_{j=1}^{\ell} n_{\cdot j} m'_j = \frac{n_{\cdot 1} m'_1 + \cdots + n_{\cdot \ell} m'_\ell}{n}. \tag{3}$$

De la même façon puisque la variance est une moyenne (à savoir celle des carrés des écarts à la moyenne), on a

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\big(m_i - \mathrm{Moy}(X)\big)^2 = \frac{n_{1\bullet}\big(m_1 - \mathrm{Moy}(X)\big)^2 + \cdots + n_{k\bullet}\big(m_k - \mathrm{Moy}(X)\big)^2}{n} \tag{4}$$

et

$$\mathrm{Var}(Y) = \frac{1}{n}\sum_{j=1}^{\ell} n_{\bullet j}\Big(m'_j - \mathrm{Moy}(Y)\Big)^2 = \frac{n_{\bullet 1}\big(m'_1 - \mathrm{Moy}(Y)\big)^2 + \cdots + n_{\bullet\ell}\big(m'_\ell - \mathrm{Moy}(Y)\big)^2}{n}. \tag{5}$$

# 4 Nuage de points

On traîte le cas de caractères *discret*. Lorsque le caractère est continu, on se ramène au cas discret en remplaçant les intervalles que sont les modalités par leurs centres.

Pour tracer le nuage de points des caractères $X$ et $Y$ dans un repère orthonormé[1], on représente chaque couple de modalités $(m_i, m'_j)$ d'effectif $n_{ij}$ non nul par un point $M_{ij}$ de coordonnées $(m_i, m'_j)$. Il faut imaginer que chaque point est muni d'un poids égal à l'effectif associé au couple de modalités qu'il représente. Le nombre total de points est donc le nombre de couples de modalités $k\ell$ et la somme des poids des points est l'effectif $n$.

On peut alors donner une interprétation géométrique de la moyenne : le centre de gravité du nuage de points est le point de coordonnées $\big(\mathrm{Moy}(X), \mathrm{Moy}(Y)\big)$. Cela signifie que si l'on imagine les points du nuage placés (avec leurs poids correspondants) sur une plaque horizontal, il suffit de placer une tige vertical sous le plateau en appui sur le point de coordonnées $\big(\mathrm{Moy}(X), \mathrm{Moy}(Y)\big)$ pour maintenir le plateau en équilibre horizontal. On donne une preuve de ce fait au paragraphe 7.1 page 6.

Chercher à *expliquer $Y$ par $X$* c'est chercher une fonction dont le graphe approche bien le nuage de points. C'est un problème compliqué[2] que l'on ne va étudier que dans un cas simple.

# 5 Régression linéaire : méthode des moindres carrés

Quand on fait de la régression linéaire, on cherche à approcher un nuage de points par le graphe d'une fonction parmi les plus simples : une droite. Les erreurs sont mesurées à l'aide des carrés des écarts verticaux entre les points du nuage et la droite (voir la figure 1 page suivante). L'erreur totale commise est la somme des carrés des écarts verticaux entre les points du nuage et la droite. En procédant de la sorte, on met en valeur les grands écarts verticaux et on dévalorise les petits écarts[3].

[1] On demande donc qu'il y ait un axe vertical et un axe horizontal et que l'echelle soit la même sur les deux axes.

[2] La définition de « approche bien » est déjà en soi un problème compliqué.

[3] On comparera avec la notion de variance. Voir le cours « Paramètres statistiques » disponible sur

http://www.univ-montp3.fr/miap/ens/AES/XA100M/index.html

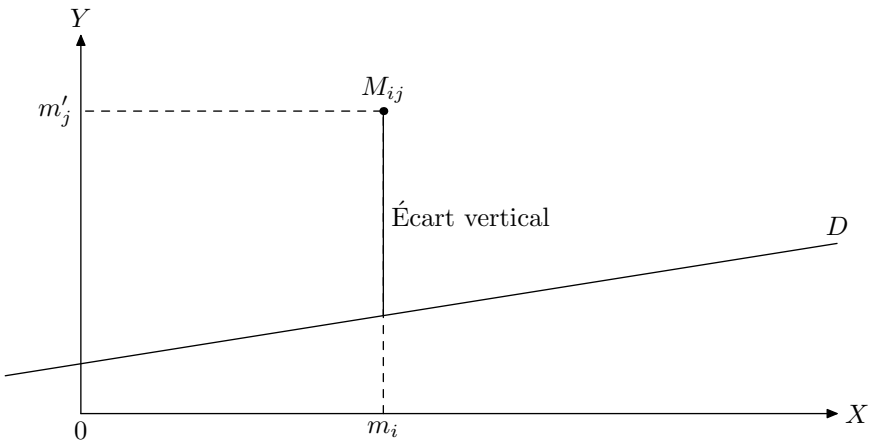


FIG. 1 – Écart vertical entre $M_{ij}$ et $D$

Pour décrire le résultat, on a besoin d'introduire la *covariance* de $X$ et $Y$ : c'est le nombre défini par

$$\mathrm{Cov}(X,Y) = \frac{1}{n} \sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}\, m_i\, m'_j - \mathrm{Moy}(X)\,\mathrm{Moy}(Y).$$

On peut retenir que c'est la moyenne des produits moins le produit des moyennes.

On montre alors (voir le paragraphe 7.2 page 7) que la droite pour laquelle l'erreur est la plus petite est la droite passant par le point de coordonnées $\big(\mathrm{Moy}(X),\mathrm{Moy}(Y)\big)$ et de coefficient directeur la covariance $\mathrm{Cov}(X,Y)$ de $X$ et $Y$ divisée par la variance $\mathrm{Var}(X)$ de $X$. C'est donc la droite d'équation

$$y = ax + b$$

avec

$$a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$

et

$$b = \mathrm{Moy}(Y) - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\,\mathrm{Moy}(X).$$

On peut aussi déterminer dans quelle mesure la droite trouvée approche bien le nuage de points. Pour cela, on introduit le *coefficient de corrélation*

$$r(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

et on montre (voir le paragraphe 7.3 page 8) que plus le coefficient $r(X,Y)$ est proche de 1 ou $-1$, meilleure est l'approximation. On verra aussi au paragraphe 7.3 que, lorsque $\mathrm{Var}(X) < \mathrm{Var}(Y)$, on commet une erreur moins grande en expliquant $X$ par $Y$ qu'en expliquant $Y$ par $X$. Pour estimer $X$ par $Y$, on utilise les mêmes formules pour $a$ et $b$ en échangeant les rôles de $X$ et $Y$.

# 6 Un exemple simple

On étudie deux caractéres $X$ et $Y$ dont le tableau de contigence est le tableau 3 page suivante.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 1 | 1 | 0 |
| 3 | 0 | 0 | 1 | 1 | 0 |
| 4 | 0 | 1 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 |

TAB. 3 –

On calcule

$$\text{Moy}(X) = 3,14 \qquad \text{Moy}(Y) = 2,86$$

$$\text{Var}(X) = 1,55 \qquad \text{Var}(Y) = 1,55$$

et

$$\text{Cov}(X, Y) = -1,41.$$

Le coefficient directeur de la droite d'ajustement est

$$a = -0,91$$

et le coefficient de corrélation est

$$r = -0,91.$$

Le nuage et la droite d'ajustement sont donnés à la figure 2.

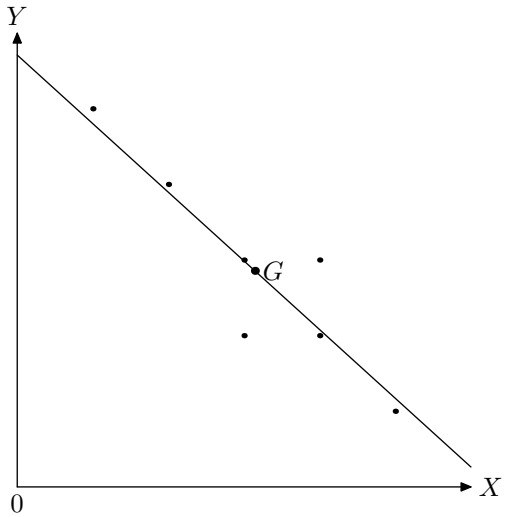


FIG. 2 – Nuage et droite d'ajustement associés au tableau 3

# 7 Annexes

## 7.1 Calcul du centre de gravité du nuage de points

Soit $G$ le centre de gravité du nuage de points. Ce nuage est composé des points $M_{ij}$ de coordonnées $(m_i, m'_j)$ et de poids $n_{ij}$. Par définition du centre de gravité $G$, on a alors

$$\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}\overrightarrow{GM_{ij}} = \vec{0}. \tag{6}$$

On note $(x_G, y_G)$ les coordonnées de $G$. La considération des abscisses dans (6) conduit à

$$\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}(m_i - x_G) = 0$$

donc

$$\begin{aligned}\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}m_i &= \sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}x_G \\ &= nx_G \text{ grâce à (1)}\end{aligned}$$

puis

$$x_G = \frac{1}{n}\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}m_i. \tag{7}$$

Mais

$$\begin{aligned}\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}m_i &= \sum_{i=1}^{k}\left(\sum_{j=1}^{\ell} n_{ij}\right)m_i \\ &= \sum_{i=1}^{k} n_{i\cdot}m_i.\end{aligned}$$

L'équation (7) devient donc

$$x_G = \frac{1}{n}\sum_{i=1}^{k} n_{i\cdot}m_i = \text{Moy}(X)$$

grâce à (2). La considération des ordonnées dans (6) conduit à

$$\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}(m'_j - y_G) = 0$$

donc

$$\begin{aligned}\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}m'_j &= \sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}y_G \\ &= ny_G \text{ grâce à (1)}\end{aligned}$$

puis

$$y_G = \frac{1}{n}\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}m'_j. \tag{8}$$

Mais

$$\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}m'_j = \sum_{j=1}^{\ell}\left(\sum_{i=1}^{k} n_{ij}\right)m'_j$$
$$= \sum_{j=1}^{\ell} n_{\cdot j}m'_j.$$

L'équation (8) devient donc

$$y_G = \frac{1}{n}\sum_{j=1}^{\ell} n_{\cdot j}m'_j = \mathrm{Moy}(Y)$$

grâce à (3).

## 7.2 Détermination de la droite d'ajustement linéaire

L'écart vertical entre le point $M_{ij}$ de coordonnées $(m_i, m'_j)$ et la droite d'équation $y = ax+b$ étant $m'_j - (am_i + b)$, on cherche $a$ et $b$ pour que la grandeur

$$T(a,b) = \frac{1}{n}\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}[m'_j - (am_i+b)]^2$$

soit minimum. En développant le carré, on obtient

$$T(a,b) = \sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}({m'_j}^2 + a^2m_i^2 + 2abm_i + b^2 - 2am_im'_j - 2bm'_j)$$

et donc

$$T(a,b) = \sum_{j=1}^{\ell} n_{\cdot j}{m'_j}^2 + a^2\sum_{i=1}^{k} n_{i\cdot}m_i^2 + 2ab\sum_{i=1}^{k} n_{i\cdot}m_i + b^2$$
$$- 2a\sum_{(i,j)\in\{1,\dots,k\}\times\{1,\dots,\ell\}} n_{ij}m_im'_j - 2b\sum_{j=1}^{\ell} n_{\cdot j}m'_j.$$

On en déduit

$$T(a,b) = \mathrm{Moy}(Y^2) + a^2[\mathrm{Var}(X) + \mathrm{Moy}(X)^2] + 2ab\,\mathrm{Moy}(X) + b^2$$
$$- 2a[\mathrm{Cov}(X,Y) + \mathrm{Moy}(X)\,\mathrm{Moy}(Y)] - 2b\,\mathrm{Moy}(Y). \quad (9)$$

Pour les valeurs de $a$ et $b$ réalisant le minimum de $T(a,b)$, on a

$$\frac{\partial T}{\partial a} = 0 \quad (10)$$

et

$$\frac{\partial T}{\partial b} = 0. \quad (11)$$

Grâce à l'équation 9, on calcule

$$\frac{\partial T}{\partial b} = 2a\,\mathrm{Moy}(X) + 2b - 2\,\mathrm{Moy}(Y)$$

et donc, l'équation 11 page précédente devient

$$a\,\mathrm{Moy}(X) + b = \mathrm{Moy}(Y). \tag{12}$$

Le point de coordonnées $\big(\mathrm{Moy}(X), \mathrm{Moy}(Y)\big)$ appartient à la droite recherchée. Grâce à l'équation 9 page précédente, on calcule

$$\frac{\partial T}{\partial a} = 2a[\mathrm{Var}(X) + \mathrm{Moy}(X)^2] + 2b\,\mathrm{Moy}(X) - 2[\mathrm{Cov}(X,Y) + \mathrm{Moy}(X)\,\mathrm{Moy}(Y)].$$

En utilisant l'équation 12, on

$$b = \mathrm{Moy}(Y) - a\,\mathrm{Moy}(X) \tag{13}$$

et donc

$$\frac{\partial T}{\partial a} = 2a[\mathrm{Var}(X) + \mathrm{Moy}(X)^2] + 2[\mathrm{Moy}(Y) - a\,\mathrm{Moy}(X)]\,\mathrm{Moy}(X) - 2[\mathrm{Cov}(X,Y) + \mathrm{Moy}(X)\,\mathrm{Moy}(Y)].$$

L'équation 10 page précédente devient alors

$$a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}. \tag{14}$$

## 7.3 Calcul d'erreur en ajustement linéaire

L'erreur commise en remplaçant le nuage de points par la droite trouvée au paragraphe précédent est $T(a,b)$ calculée en 9 page précédente. En remplaçant $a$ et $b$ par leurs valeurs trouvée en 14 et 13 on obtient

$$T(a,b) = \mathrm{Var}(Y)\left[1 - \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(X)\,\mathrm{Var}(Y)}\right].$$