Université Paris Diderot **Économétrie**
Licence 3, 2012-2013 **Vianney Perchet - Aurélie Fischer**

T.D. 3 : Modèle linéaire multiple

Ce T.D. est consacré au modèle linéaire multiple. Le cas particulier du modèle linéaire simple est étudié dans l'exercice 1. Un exemple d'application du modèle linéaire multiple est proposé dans l'exercice 2. Dans les exercices 3 et 4, on se pose la question suivante : que se passe-t-il si l'on oublie d'inclure une variable dans la régression ? L'exercice 3 présente une étude réelle, menée par deux économistes, sur l'équilibre entre temps de travail et temps de sommeil.

Exercice 1. Soit le modèle linéaire gaussien simple :

$$Y_i = \beta + \alpha x_i + \varepsilon_i, \quad \text{avec } i = 1, 2, \ldots, n\,.$$

1) Ecrire sous forme matricielle le modèle linéaire gaussien simple.
2) Expliciter les matrices ${}^tXX$ et ${}^tXY$.
3) Expliciter la matrice $({}^tXX)^{-1}$.
4) Expliciter l'estimateur des moindres carrés $T = {}^t(B, A)$ du vecteur des paramètres $\theta = {}^t(\beta, \alpha)$.
5) Expliciter la loi de T. Retrouver alors la loi de A et de B.
6) Soit $w = {}^t(w_1, w_2) \in \mathbb{R}^2$. Expliciter l'intervalle de confiance du paramètre ${}^tw\theta$.
7) En choisissant judicieusement w, retrouver l'expression des intervalles de confiance des paramètres du modèle linéaire gaussien simple.

Exercice 2. Chez des patients ayant des problèmes cardiaques, on a mesuré (par effet Doppler) la vitesse de circulation du sang Y dans les artères coronaires. On cherche à étudier l'effet de deux variables quantitatives sur cette vitesse, à savoir le taux de cholestérol t et le poids p. On dispose des données suivantes : pour chaque patient i, $i = 1, \ldots, 20$ on mesure son poids p_i, son taux de cholestérol t_i, et sa vitesse de circulation sanguine y_i.

p_i	t_i	y_i	p_i	t_i	y_i	p_i	t_i	y_i
45	2,7	75,09	58	2,1	72,42	79	3,0	64,22
48	2,0	77,41	63	1,8	69,63	79	1,9	66,34
50	1,8	77,88	66	2,4	70,60	84	2,1	62,34
50	2,2	76,52	66	2,9	68,62	89	1,8	61,06
52	1,7	77,00	69	2,0	68,53	90	2,5	59,68
53	2,5	72,09	72	2,6	67,88	98	2,9	55,81
56	2,8	71,96	74	1,7	66,28			

On donne

$$\begin{array}{lll} \sum_{i=1}^{20} p_i = 1341 & \sum_{i=1}^{20} p_i^2 = 94607 & \sum_{i=1}^{20} p_i y_i = 90791,2 \\ \sum_{i=1}^{20} t_i = 45,4 & \sum_{i=1}^{20} t_i^2 = 106,74 & \sum_{i=1}^{20} t_i y_i = 3121,36 \\ \sum_{i=1}^{20} y_i = 1381,36 & \sum_{i=1}^{20} y_i^2 = 96155,04 & \sum_{i=1}^{20} p_i t_i = 3064,1\,. \end{array}$$

1) Rappeler la modélisation du modèle linéaire gaussien (notation en équation et matricielle).

2) Donner les estimations des paramètres β, α_1 et α_2. On donne :

$$\begin{pmatrix} 20 & 1341 & 45,4 \\ 1341 & 94607 & 3064,1 \\ 45,4 & 3064,1 & 106,74 \end{pmatrix}^{-1} = 10^{-3} \begin{pmatrix} 2102,24 & -11,9331 & -551,5969 \\ -11,9331 & 0,21815 & -1,1867 \\ -551,5969 & -1,1867 & 278,0473 \end{pmatrix}.$$

3) Donner des intervalles de confiance, au niveau de confiance 95% des paramètres β, α_1 et α_2 sachant que $\sum_{i=1}^{20}(y_i - b - a_1 p_i - a_2 t_i)^2 = 22,631$.

Exercice 3. On s'intéresse au modèle linéaire suivant :

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + \varepsilon_i.$$

On suppose que les variables explicatives sont orthogonales, ce qui signifie que pour tout $(j,k) \in \{1,\dots,p\}$, où $j \neq k$,

$$\sum_{i=1}^{n} x_{ij} x_{ik} = 0,$$

et qu'elles sont centrées, c'est-à-dire, pour tout $j = 1,\dots,p$,

$$\sum_{i=1}^{n} x_{ij} = 0.$$

1) Ecrire le modèle sous forme matricielle.
2) Appliquer la formule du cours pour calculer l'estimateur des moindres carrés $\hat{b}$, et vérifier que dans ce cas, on obtient une formule simple pour chacune des coordonnées de $\hat{b}$, que l'on notera $\hat{b}_j$ pour $j = 0, ..., p$.
3) Sous l'hypothèse que les ε_i sont i.i.d. $\mathcal{N}(0,\sigma^2)$, calculer la loi de chaque $\hat{b}_j$.
4) Un économètre tente de modéliser la variable y en fonction d'autres variables. Malheureusement, il ne dispose pas de toute l'information nécessaire et commet une erreur dans son modèle en omettant une variable, par exemple x_p, c'est-à-dire qu'il estime le modèle correspondant à
$$y_i \simeq \sum_{j=1}^{p-1} b_j x_{ij}.$$
D'après la question (2), qu'obtient-il alors comme estimateur des moindres carrés ? Que vaut l'espérance de cet estimateur ?
5) Conclure : dans le cas d'orthogonalité des variables, quelle est la conséquence de l'oubli d'une variable sur l'estimation de l'effet des autres variables ?

Exercice 4. Un criminologue essaie de mesurer l'impact du taux de chômage t_i et du nombre de policiers pour 10000 habitants x_i sur le taux de criminalité y_i (nombre de crimes et délits constatés pour 10000 habitants), dans n départements $i = 1, ..., n$. Il propose le modèle

$$y_i = b_0 + b_1 x_i + b_2 t_i + \varepsilon_i \tag{1}$$

et obtient

$$\widehat{y_i} = 10,52 - 0,08 x_i + 1,27 t_i.$$

Un autre criminologue, niant un effet du chômage sur la criminalité, postule le modèle

$$y_i = c_0 + c_1 x_i + \varepsilon_i$$

et obtient, avec les mêmes données,

$$\widehat{y}_i = 11,83 + 0,87 x_i.$$

1) En quoi ceci semble-t-il paradoxal ?
2) Montrer que si le modèle donné par l'équation (1) est correct, on a :

$$\mathbb{E}(\hat{c}_1) = b_1 + \delta b_2,$$

où δ est le coefficient de la régression de t_i sur x_i.

3) En déduire le signe vraisemblable de δ.
4) Commenter : quelles sont les raisons d'un tel phénomène ?
5) Conclure pour les deux exercices (1 et 2) : que penser du problème de l'omission d'une variable pertinente en général ?

Exercice 5. Cet exemple est tiré du livre de Wooldridge. Deux économistes américains, Biddle et Hammermesh, ont étudié le temps de sommeil y_i d'adultes $i = 1, ..., n$ en fonction de leur temps de travail par semaine t_i (les deux durées mesurées en minutes/semaine), de leur niveau d'études en années, e_i et de leur âge a_i. Ils souhaitaient étudier un effet de balance entre temps de travail et temps de sommeil (en gros, est-ce qu'un travailleur est prêt à dormir moins pour travailler plus et donc gagner plus, ou à l'inverse à dormir plus quitte à travailler moins et gagner moins). Le modèle est le suivant :

$$y_i = b_0 + b_1 t_i + b_2 e_i + b_3 a_i + \varepsilon_i.$$

1) Si l'effet de balance existe, quel est le signe attendu de b_1 ?
2) Le modèle estimé obtenu est :

$$\widehat{y}_i = 3638,25 - 0,148 t_i - 11,13 e_i + 2,20 a_i + \varepsilon_i.$$

D'après ce modèle, si un adulte travaille 5 heures de plus par semaine, combien de temps dormira-t-il en moins en moyenne ? Commenter cette valeur.

3) Que penser du signe et de la valeur du coefficient correspondant au niveau d'études ?
4) Pensez-vous que le temps de travail, le niveau d'études et l'âge soient des variables pertinentes pour expliquer la variable "sommeil" ? Quels autres facteurs pourraient contribuer à expliquer y ?