Université Paris 7 **Économétrie**
Licence 3, 2009-2010 **Pierre Alquier - Joseph Salmon**

TD N° 3 : Modèle linéaire multiple - exemples

Dans ce TD, on fait des manipulations de modèles linéaires. Dans les exercices 1 et 2, on se pose la question suivante : que se passe-t'il si on oublie d'inclure une variable dans la régression ? Le troisième présente une étude réelle, menée par deux économistes, sur l'arbitrage entre temps de travail et temps de sommeil.

Exercice 1. On s'intéresse au modèle linéaire suivant :

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

dans le cas très particulier où toutes les variables sont orthgonales entre elles et centrées :

$$\sum_{i=1}^{n} x_{j,i} x_{k,i} = 0 \text{ si } k \neq j, \text{ et } \sum_{j=1}^{n} x_{j,i} = 0.$$

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 $j = 1, ..., p$.
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) On suppose qu'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
$$y_i \simeq \sum_{j=1}^{p-1} b_j x_{j,i}.$$
D'après la question (2), quelle sera la valeur de l'estimateur des moindres carrés qu'il obtiendra ici ?
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 2. Un criminologue essaie de mesurer l'impact du taux de chomage, chom_i et du nombre de policiers pour 10000 habitants, police_i dans n départements $i = 1, ..., n$, sur le taux de criminalité crim_i pour 10000 habitants. Il propose le modèle suivant :

$$\text{crim}_i = b_0 + b_1 \text{police}_i + b_2 \text{chom}_i + \varepsilon_i \tag{1}$$

et construit les matrices et vecteurs

$$\text{crim} = \begin{pmatrix} \text{crim}_1 \\ \vdots \\ \text{crim}_n \end{pmatrix} \text{ et } X = \begin{pmatrix} 1 & \text{police}_1 & \text{chom}_1 \\ \vdots & \vdots & \vdots \\ 1 & \text{police}_n & \text{chom}_n \end{pmatrix}.$$

Il obtient le modèle estimé suivant

$$\widehat{\text{crim}}_i = 10.52 - 0.08\text{police}_i + 1.27\text{chom}_i.$$

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

$$\text{crim}_i = c_0 + c_1\text{police}_i + \varepsilon_i$$

et obtient, **avec les mêmes données**, le modèle estimé suivant :

$$\widehat{\text{crim}}_i = 11.83 + 0.87\text{police}_i.$$

1) En quoi ceci semble-t'il paradoxal ?
2) Montrer que, si c'est le modèle donné par l'Equation (1) qui est correct, on a alors la formule suivante :
$$\mathbb{E}(\hat{c}_1) = b_1 + \delta b_2$$
où δ est le coefficient de la régression de police_i sur chom_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 3. Cet exemple est tiré du livre de Wooldridge. Deux économistes américains, Biddle et Hammermesh, ont étudié le temps de sommeil sommeil_i d'adultes $i = 1, ..., n$ en fonction de leur temps de travail par semaine, travail_i (les deux temps mesurés en minutes/semaine), de leur niveau d'études en années, etudes_i, et de leur âge, age_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 dormir moins). Le modèle est le suivant :

$$\text{sommeil}_i = b_0 + b_1\text{travail}_i + b_2\text{etudes}_i + b_3\text{age}_i + \varepsilon_i.$$

1) Si l'effet de balance existe, quel est le signe attendu de b_1 ?
2) Que peut-on attendre comme signes pour b_2 et b_3 ?
3) Le modèle estimé obtenu est :
$$\widehat{\text{sommeil}}_i = 3638.25 - 0.148\text{travail}_i - 11.13\text{etudes}_i + 2.20\text{age}_i + \varepsilon_i.$$
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.
4) Que penser du signe et de la valeur du coefficient correspondant aux études ?
5) Pensez-vous que travail_i, etudes_i et age_i sont des variables pertinentes pour expliquer la variable sommeil_i ? Quels autres facteurs pourraient contribuer à expliquer sommeil_i ?