STATISTIQUE - Econométrie

Jacqueline PRADEL
pradelj@univ-paris1.fr
courrier : MSE
Licence Sciences Economiques
UFR 02, Université Paris 1
http://econometrie-mse.univ-paris1.fr

1

Les étudiants sont invités à transmettre toutes leurs questions et remarques par le courrier interne (déposé à la loge du Panthéon, avec simple mention du centre « MSE »), ainsi que par e-mail pour ceux qui disposent d 'un accès internet.

Le cours est pour la plupart d'entre les étudiants le dernier cours de stat qu'ils auront. Les notions qui vont être étudiées s'appuient sur les cours de DEUG de proba et stat, et seront systématiquement introduites en vue de leurs applications en économie, en évitant le plus possible les calculs compliqués, mais pas les raisonnements complexes !

Menu

I. Introduction

modèles statistiques

II. Estimation, prévision

- *ponctuelles*
- *par intervalle*

III. TESTS

Application au
Modèle Statistique Linéaire
$y = ax + b + u$

2

Vous avez déjà utilisé des modèles statistiques sans le savoir, nous allons préciser et étendre le concept

Vous avez étudié l'estimation ponctuelle d'un paramètre à partir d'un échantillon. Nous allons réviser ces notions et les compléter : à la question « *quelle est la valeur de* θ *?* » on ajoute la question « *dans quelle fourchette se trouve* θ *?* ». Cela conduit à l 'estimation par intervalle.

Puis on peut se poser le *même* type de *questions pour y et non pour* θ . Cela conduit aux prévisions, ponctuelles ou par intervalle

Enfin, on peut avoir à répondre à une question du genre « θ *vérifie-t-il telle condition, ou non?* ». C'est le problème des tests : chercher la réponse, c'est tester la condition en question.

Toutes ces notions seront appliquées notamment au cas du modèle statistique linéaire standard « y = ax + b + u », avec les estimations, prévisions et tests liés à la méthode des moindres carrés ordinaires : c'est la base élémentaire de l'économétrie. Il convient d'en connaître le bon usage et les limites.

Bibliographie pour mise à niveau :

Statistique Probabilités, Estimation Ponctuelle : Bouzitat-Bouzitat-Pagès, Cujas 90

I.1 Modèles Statistiques

Modèle : relation ± simplifiée entre les variables

Exemple $y = a\,x + b$

Modèle statistique : $(\mathcal{Y}, \mathcal{P})$ ou $(\mathcal{Y}, \mathcal{X}, \mathcal{P})$

$\mathcal{Y}$: ensemble des valeurs R^N des variables expliquées

$\mathcal{X}$: ensemble des valeurs R^{KN} des variables explicatives

$\mathcal{P}$: ensemble des lois de probabilité des variables.

Exemple $y_n = a\,x_n + b + u_n \qquad n = 1,\dots,N$

$E(u_n) = 0$

3

Un modèle théorique est une équation reliant une variable « expliquée » à une ou des variables « explicatives ».. Il est vrai qu'il n'est pas nécessaire que la relation écrite soit « la vraie relation » (heureusement!) : de même que l'image d'un « photon » qui « rebondit » sur un miroir peut permettre de construire un périscope bien qu'elle ne corresponde pas à la réalité physique, de même un modèle correctement construit pourra être utile, au moins « localement », c 'est-à-dire toutes choses restant comparables. Par contre, c'est une réflexion théorique sur le phénomène étudié qui permet de choisir les variables « explicatives », c'est-à-dire dont les variations peuvent expliquer celles de la variable que l'on cherche à expliquer, et la forme de cette relation.

On espère ainsi pouvoir répondre à des questions du type :

• Quel sera l'impact sur les ventes de tel produit si j'augmente son prix de 1% ? On sent bien que l'équation envisagée sera différente selon le pouvoir de monopole que je détiens sur le marché de ce produit et la présence de produits substituables sur le marché.

• Quels sont les déterminants du prix d'un appartement ? On fera intervenir sa surface, le nombre de pièces, l'étage, l'âge de l'immeuble, sa situation géographique, etc…

• Comment l'activité innovatrice d'une entreprise est-elle liée aux performances de cette entreprise ? On fera intervenir le type d'innovation (produit ou procédé), son intensité (amélioration d'une technique ou d'un produit existant, ou première technologique), etc..

Le modèle statistique correspondant sera obtenu en introduisant des fluctuations aléatoires autour du modèle théorique pour tenir compte des facteurs non pris en compte dans le modèle (hétérogénéité non observée sur les individus).

Modèles Statistiques avec Paramètre

Soit un modèle $(\mathcal{Y}, \mathcal{P}, \Theta)$ ou $(\mathcal{Y}, \mathcal{X}, \mathcal{P}, \Theta)$

- **Modèle paramétrique**

  à θ donné est associée une seule loi dans $\mathcal{P}$

- **Modèle semi-paramétrique**

  à θ donné sont associées plusieurs lois dans $\mathcal{P}$

- **Paramètre identifiable**

  Chaque loi dans $\mathcal{P}$ n'est associée qu'à un seul θ

- **Paramètre non identifiable**

  La même loi dans $\mathcal{P}$ peut être associée à plusieurs θ

4

Exemple Variables observées = y_n , x_n pour n = 1,…, N

$\mathcal{Y} = R^N$ et $\mathcal{X} = R^N$

Ensemble des lois de probabilité des variables défini par la propriété :

$$E(y_n) = a\, x_n + b, \text{ où a et b sont deux réels.}$$

Nous avons introduit deux paramètres, a et b.

$$\theta = (a, b) \in \Theta = R^2$$

(i) La loi des observations est-elle entièrement déterminée lorsque θ est donné ?

On sait que les x_n sont non aléatoires (certaines), mais pour les y_n, on ne connaît que leur espérance lorsque a et b sont connus : pour un θ donné, il y a de nombreuses lois possibles. Le modèle est semi-paramétrique

Si on rajoute : $y_n \sim N(a\, x_n + b, \sigma^2)$ indépendantes, alors la loi de $(y_1, y_2, \ldots, y_N)$ est entièrement connue dès que θ est donné : le modèle devient un modèle paramétrique.

(ii) la valeur de θ est-elle entièrement déterminée lorsque la loi est donnée ?

$\Leftrightarrow$ le paramètre est-il identifiable ?

L'espérance est entièrement déterminée, mais cela ne définit a et b de manière unique que si il existe au moins deux valeurs différentes pour les x_n

Si x_n pour tout n, alors $E(y_n) = ax + b$ peut aussi s 'écrire

$$E(y_n) = ax + b = a^*x + (b + ax - a^*x) = a^*x + b^*$$

où b^* est également une constante : le couple (a^*, b^*) peut être associé à la même loi que (a, b) : le paramètre θ n'est pas identifiable.

S'il existe deux valeurs différentes, par exemple $x_1 \neq x_2$, a et b sont définis par les deux équations $E(y_1) = ax_1 + b$ et $E(y_2) = ax_2 + b$.

I.2 Inférence Statistique

- Estimation : quelle(s) valeur(s) de θ ?
- Test : θ vérifie-t-il l'hypothèse H_o, ou non ?
- Prévision : quelle(s) valeur(s) pour y ?
- Simulation : tirage au hasard de valeurs de y selon le système générateur des données (DGP in english).

5

Estimation :	une valeur = estimation ponctuelle = rappels et compléments
	un intervalle = estimation par intervalle
Tests :	de spécification (mon modèle est-il acceptable?)
	de contraintes (pourrais-je avoir moins de paramètres?)
Prévision :	mêmes questions que pour θ, mais concernant y
Simulations :	ne sera pas traité dans ce cours

I.3 Exemples de Modèles Statistiques

- Modèle d'échantillonnage

  $(y_1, y_2, \ldots, y_N)$ i.i.d. selon loi $L(y;\theta)$, $\theta \in \Theta$

- Modèle linéaire

  Modèle linéaire simple

  Modèle linéaire standard normal

6

Les échantillons sont les seuls modèles rencontrés jusqu'à présent :

$(y_1, y_2, \ldots y_N)$ indépendants et de même loi L forment un n-échantillon (ou échantillon de taille N) de la variable « parente » de loi L

vous verrez en TD les échantillons de variable de Bernoulli, de variable de Poisson, de variable exponentielle, de variable uniforme sur un intervalle et de variable normale.

La nouveauté de cette année : introduction de variables explicatives x, et modélisation de la loi de y sachant x (c'est-à-dire que l'on traite x comme étant fixée, non aléatoire).

X non constante.

Modèle linéaire simple :	H1	$E(y_n \mid x_n) = a \ x_n + b$
Modèle linéaire standard :	H1+H2	$V(y_n \mid x_n) = \sigma^2$ et les y_n sont non corrélées
Modèle linéaire standard normal :		$y_n \sim N(ax_n + b, \sigma^2)$ indépendantes

Les modèles à variables latentes ne sont pas au programme de ce cours, mais font l'objet d'exercices d'approfondissement pour les étudiants motivés. Dans ces modèles, la variable observée est tronquée ou censurée, ou même uniquement positionnée par rapport à un seuil.