TD Econométrie 2

Introduction

Dans le cours deux sources d'endogénéité des variables explicatives ont été présentées en détail : la simultanéité des variables explicatives avec la variable à expliquer (e.g., cas des données de marché : les prix et les quantités offertes et/ou demandées sont des variables simultanées en raison du mécanisme d'ajustement des prix nécessaire à l'équilibre du marché) et les erreurs de mesure sur les variables explicatives. Le principal objectif de ce TD est de présenter la troisième source d'endogénéité des variables explicatives d'un modèle : l'omission de variables explicatives pertinentes.
Le problème de l'omission de variables explicatives est très important en économétrie. Ce problème survient généralement lorsque le modèle utilisé est mal spécifié, i.e. que certains effets ont été oubliés dans le modèle utilisé. Parmi les effets les plus fréquemment omis, on peut citer les effets croisés des variables explicatives et les effets de variables qui sans être intéressants en tant que tels interagissent tout de même avec la relation étudiée. Par exemple, dans un modèle de demande de marché si on oublie de tenir compte de l'évolution de la population et de sa structure socio-démographique, les élasticités-prix calculées peuvent être complètement (asymptotiquement) biaisées.
Par ailleurs, les économètres utilisent souvent des données qu'ils ne produisent pas. Aussi, souvent ils aimeraient utiliser dans leurs modèles des variables qui ne sont pas mesurées. En ce sens, ils n'omettent pas vraiment ces variables, ils ne peuvent tout simplement pas utiliser des variables qui seraient effectivement pertinentes mais qui ne sont pas mesurées (voire mesurables).
En pratique les conséquences en matière d'estimation de cette non-utilisation de variables explicatives pertinentes peuvent être importantes. L'objet de la première partie de l'exercice est d'illustrer ce problème. Le but de la fin de l'exercice est de décrire les deux principales méthodes utilisables pour contourner ce problème. Les principales solutions au problème de l'omission des variables explicatives sont le recours à des variables de contrôle (de l'hétérogénéité des relations étudiées), ou tout au moins à des proxies, ou le recours à des variables instrumentales.

Omission de variables explicatives pertinentes et endogénéité : cas des équations de salaire

Nous prendrons l'exemple, assez emblématique, de la mesure de l'effet des années d'étude sur le salaire des jeunes (après cinq ans d'emploi, pour des hommes ayant au moins le baccalauréat et travaillant dans le même secteur d'activité, pour avoir une relation homogène). Supposons que le modèle économétrique donné par les équations:

$$y_i = \alpha + \beta n_i + \delta q_i + v_i \quad \text{et} \quad E[v_i / n_i, q_i] = E[v_i] = 0$$

soit correctement spécifié. Ici le paramètre d'intérêt est le paramètre β celui qui lie n_i : le nombre de semestres d'études à partir du baccalauréat à y_i le salaire des jeunes, les variables économiques d'intérêt. La variable q_i n'est qu'une variable décrivant ce qui peut affecter le salaire en dehors des années d'études, disons l'« agilité intellectuelle » du salarié, celle qui lui

permet de s'adapter rapidement et de répondre rapidement aux demandes qui lui sont adressées.[1] Un échantillon de N (avec N grand) observations de (y_i, n_i, q_i) i.i.d. est disponible. Nous utiliserons les notations suivantes :

$$\mathbf{x}_i \equiv \begin{bmatrix} 1 \\ n_i \end{bmatrix}, \ \mathbf{x}_{qi} \equiv \begin{bmatrix} 1 \\ n_i \\ q_i \end{bmatrix}, \ \mathbf{a}_{q0} \equiv \begin{bmatrix} \alpha \\ \beta \\ \delta \end{bmatrix}$$

et les notations habituelles pour l' « empilement » des variables dans des matrices correspondant aux données pour l'ensemble de l'échantillon.

1. Donner l'interprétation des termes α, β, δ **et** v_i

2. Le cas simple : q_i **est mesurée**

2.1. Montrer que si q_i ***était mesurée, il serait possible de construire un estimateur des paramètres d'intérêt du modèle***

2.2. Donner les propriétés de cet estimateur et sa distribution

3. Conséquences de l'omission de q_i

Ici nous supposons que q_i n'est pas disponible dans l'échantillon, ce qui est généralement le cas. On se propose alors de travailler sur un modèle contenant β, notre paramètre d'intérêt, de la forme suivante :

$$y_i = \lambda + \beta n_i + u_i \quad \text{avec} \quad E[u_i] = 0$$

ou, en version « compacte » :

$$y_i = \mathbf{a}_0' \mathbf{x}_i + u_i \ \text{avec} \ E[u_i] = 0 \ \text{où} \ \mathbf{a}_0 \equiv \begin{bmatrix} \lambda \\ \beta \end{bmatrix}.$$

Dans la suite $\overline{x}$ représentera la moyenne empirique des n_i observées :

$$\overline{n} \equiv \frac{1}{N} \sum_{i=1}^{N} n_i$$

et $\overline{y}$ celle des y_i.

3.1. Donner l'interprétation des termes β, λ **et** u_i, ***et montrer que ce modèle est un modèle en « moyenne »***

3.2. Donner l'expression de l'estimateur des paramètres du modèle observable par les MCO et donner les conditions sous lesquels il est convergent et les conditions sous lesquelles il ne l'est pas

[1] En pratique, on considère également une autre variable inobservée importante (latente). Il s'agit de ce qu'on peut nommer l' « ardeur au travail » du salarié. Cette variable non mesurée pose le même type de problèmes que l'ardeur au travail en matière d'inférence. Nous ignorons cet effet ici par souci de simplicité.

3.3. Montrer que la non convergence de $\hat{\beta}_N^{MCO}$ est liée au fait que n_i est endogène par rapport à u_i

3.4. Argumenter le fait que l'estimateur des paramètres du modèle observable par les MCO n'est vraisemblablement pas convergent ici et que le biais associé à l'utilisation des MCO est vraisemblablement positif

3.5. Commenter le biais lié à l'utilisation de l'estimateur des paramètres du modèle observable par les MCO lorsque les conditions de sa convergence ne sont pas satisfaites

4. Première solution à l'omission ou à la non-observation de q_i dans les équations de salaire : l'utilisation de variables instrumentales

On suppose donc ici, qu'on travaille avec le modèle observable :

$$y_i = \lambda + \beta n_i + u_i \text{ avec } E[u_i] = 0$$

où : $\lambda \equiv \alpha + \delta E[q_i]$ et $u_i = v_i + \delta(q_i - E[q_i])$. On vient de montrer que δ et $Cov[n_i, q_i]$ sont positifs selon toute vraisemblance.

4.1. Donner les propriétés de « bonnes » variables instrumentales pour ce modèle

4.2. Déterminer au moins deux variables instrumentales potentiellement utilisables pour l'équation de salaire observable

4.3. Donner un estimateur convergent des paramètres de l'équation de salaire observable en supposant que les variables instrumentales définies dans la question précédente sont mesurées.

4.4. Prouver sa convergence et donner sa distribution as.

4.5. Montrer que s'il n'y avait qu'une variable instrumentale disponible (e.g., z_{1i}), l'estimateur des 2MC se réduirait à l'estimateur des VI

5. Seconde solution à l'omission ou à la non-observation de q_i dans une équation de salaire : le contrôle de l'hétérogénéité des q_i

En repartant de l'équation de salaire initiale (structurel latente) :

$$y_i = \alpha + \beta n_i + \delta q_i + v_i \text{ et } E[v_i / n_i, q_i] = E[v_i] = 0$$

il est possible de proposer une autre solution au problème de la non-mesure de q_i. Nous supposerons que si q_i n'est pas mesurée, il existe (au moins) une variable c_i telle que :

$$q_i = bc_i + e_i \text{ avec } E[e_i / n_i, c_i] = E[e_i] = 0,\ b \neq 0 \text{ et } E[v_i / n_i, c_i] = E[v_i] = 0$$

L'objectif de cette partie est de montrer que l'utilisation de c_i, une variable de contrôle de l'hétérogénéité de q_i, peut également permettre de gérer le problème de la non-mesure de q_i.

5.1. Montrer qu'en combinant l'équation de salaire avec l'équation liant q_i à c_i il est possible d'estimer simplement le paramètre β

5.2. Interpréter l'équation $q_i = bc_i + e_i$ avec $E[e_i] = 0$ et $b \neq 0$, $E[v_i / n_i, c_i] = E[v_i] = 0$

5.3. Interpréter le terme e_i ***et la condition*** $E[e_i / n_i, c_i] = E[e_i] = 0$. ***Montrer que selon cette condition*** n_i ***n'est lié à*** q_i ***qu'à travers*** c_i.

5.4. Déterminer au moins une variable qui pourrait jouer le rôle de c_i ***dans l'équation de salaire considérée ici***