TD Econométrie 2, éléments de corrigé

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 importants. 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:

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

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) 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

- Le terme β mesure l'effet marginal (moyen) d'un semestre d'études après le bac sur le salaire :

$$\beta = \frac{\partial y_i}{\partial n_i}.$$

C'est ce qu'amène en moyenne en matière de salaire après cinq ans de travail salarié un semestre d'études après le bac. Ce effet marginal (ou partiel) est supposé constant, quelque soit le nombre d'années d'études, la filière choisie et le type de d'école choisi. Le terme δn mesure l'effet sur le salaire de n semestres d'études, c'est l'effet causal des études sur le salaire. Il mesure l'effet des acquis de connaissances liés aux études sur le salaire obtenu.

[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.

- De la même manière δ mesure l'effet marginal (moyen) de l'agilité intellectuelle du salarié sur le salaire qu'il obtient après cinq ans de salariat. Par convention nous supposerons que $q = 0$ est le niveau minimum de l'agilité intellectuelle des salariés les moins doués de l'échantillon. En fait, nous pouvons choisir cette normalisation comme on le souhaite. La q_i n'étant pas mesurée, tout ce qui concerne son niveau ou son échelle reste à un niveau théorique.

- Le terme $\alpha + v_i$ représente donc le niveau de salaire qu'aurait obtenu le salarié *i* s'il avait arrêté ses études après le bac et s'il avait l'agilité intellectuelle minimum.

- Puisque $E[v_i] = 0$, le terme α mesure le salaire moyen du salarié le moins doué et s'il a arrêté ses études après le bac. C'est le salaire de base moyen d'un salarié après cinq ans de salariat.

- Le terme v_i regroupe tout ce qui conduit au salaire y_i en dehors des effets causaux des études et de l'agilité intellectuelle, et du salaire de base moyen α.

- Nous interpréterons plus bas la condition d'exogénéité $E[v_i / n_i, q_i] = E[v_i] = 0$.

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

- Si q_i était observée, le modèle :

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

ou, en version « compacte » :

$$y_i = \mathbf{a}'_{q0}\mathbf{x}_{qi} + v_i \text{ avec } E[v_i / \mathbf{x}_{qi}] = E[v_i] = 0$$

serait un simple modèle de régression puisque toutes les variables explicatives du modèle sont exogènes et donc non corrélées au terme d'erreur :

$$E\left[\begin{bmatrix} 1 \\ n_i \\ q_i \end{bmatrix} v_i\right] = E[\mathbf{x}_{qi} v_i] = \mathbf{0}_{3\times 1}.$$

2.2. Donner les propriétés de cet estimateur et sa distribution

- Donc, dans ce cas il est possible d'utiliser un estimateur des MCO pour estimer $\mathbf{a}_{q0}$ de façon convergente :

$$\hat{\mathbf{a}}_{qN}^{MCO} = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{qi}\mathbf{x}_{qi}'\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{qi}y_i\right] \xrightarrow[N\to+\infty]{p} \mathbf{a}_{q0},$$

cet estimateur est même sans biais.

- Puisque nous n'avons *a priori* aucune indication sur la loi des v_i (sauf que les v_i sont indépendants), nous sommes contraints de « seulement » déterminer la loi as. de $\hat{\mathbf{a}}_{qN}^{MCO}$, loi qui pourra être utilisée puisque N est grand.

- D'après le cours (et le TD1), on sait que :

$$\sqrt{N}\left(\hat{\mathbf{a}}_{N}^{MCO} - \mathbf{a}_0\right) \xrightarrow[N\to+\infty]{L} \mathbf{N}\left(\mathbf{0}_{P\times 1}, \mathbf{\Sigma}_0\right)$$

avec :

$$\mathbf{\Sigma}_0 = E\left[\mathbf{x}_{qi}\mathbf{x}_{qi}'\right]^{-1}\mathbf{W}_0 E\left[\mathbf{x}_{qi}\mathbf{x}_{qi}'\right]^{-1} \text{ et } \mathbf{W}_0 \equiv V\left[\mathbf{x}_{qi}v_i\right] = E\left[\mathbf{x}_{qi}(v_i)^2\mathbf{x}_{qi}'\right].$$

- Ici aucune hypothèse n'est posée à propos de la variance des v_i, aussi il est impossible de simplifier l'expression de $\mathbf{W}_0$ grâce à l'homoscédasticité supposée des v_i (ce qui serait vraisemblablement une erreur avec ce type de données, les salaires étant très variables). Il est donc nécessaire d'utiliser l'estimateur de White (robuste à l'hétéroscédasticité) pour $\mathbf{\Sigma}_0$. La construction de cette estimateur est relativement simple, elle consiste à calculer la contre-partie empirique de $\mathbf{\Sigma}_0$, élément par élément, et de remplacer $\mathbf{a}_{q0}$ par son estimateur : $\hat{\mathbf{a}}_{qN}^{MCO}$:

$$\hat{\mathbf{\Sigma}}_N = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{qi}\mathbf{x}_{qi}'\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{qi}(v_i(\hat{\mathbf{a}}_{qN}^{MCO}))^2\mathbf{x}_{qi}'\right]\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{qi}\mathbf{x}_{qi}'\right]^{-1}$$

où :

$$v_i(\hat{\mathbf{a}}_{qN}^{MCO}) = y_i - \mathbf{x}_{qi}'\hat{\mathbf{a}}_{qN}^{MCO}$$

Puisque, $\hat{\mathbf{a}}_{qN}^{MCO} \xrightarrow[N\to+\infty]{p} \mathbf{a}_{q0}$, $v_i(\hat{\mathbf{a}}_{qN}^{MCO})$ est un estimateur convergent de $y_i - \mathbf{x}_{qi}'\mathbf{a}_{q0}$ donc de $v_i = y_i - \mathbf{x}_{qi}'\mathbf{a}_{q0}$.

- Sous certaines conditions de régularité généralement vérifiées, la *prop. A14* s'applique et on a :

$$\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{qi}(v_i(\hat{\mathbf{a}}_{qN}^{MCO}))^2\mathbf{x}'_{qi}\right]\xrightarrow[N\to+\infty]{p}\mathbf{W}_0\equiv E\left[\mathbf{x}_{qi}(v_i)^2\mathbf{x}'_{qi}\right].$$

- De même, en utilisant la loi (faible) des grands nombre et la *prop. A3*, il est possible de montrer que :

$$\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{qi}\mathbf{x}'_{qi}\right]^{-1}\xrightarrow[N\to+\infty]{p}E\left[\mathbf{x}_{qi}\mathbf{x}'_{qi}\right]^{-1}.$$

- Finalement, par la *prop. A7* de préservation de la convergence en probabilité des transformations continues des suites de var, on montre que la limite du produit des var constituant $\hat{\mathbf{\Sigma}}_N$ converge en probabilité vers le produit de leurs limites :

$$\hat{\mathbf{\Sigma}}_N=\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{qi}\mathbf{x}'_{qi}\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{qi}(v_i(\hat{\mathbf{a}}_{qN}^{MCO}))^2\mathbf{x}'_{qi}\right]\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{qi}\mathbf{x}'_{qi}\right]^{-1}$$
$$\xrightarrow[N\to+\infty]{p}E\left[\mathbf{x}_{qi}\mathbf{x}'_{qi}\right]^{-1}E\left[\mathbf{x}_{qi}(v_i)^2\mathbf{x}'_{qi}\right]E\left[\mathbf{x}_{qi}\mathbf{x}'_{qi}\right]^{-1}=\mathbf{\Sigma}_0$$

- Aussi, lorsque N est grand on a :

$$\hat{\mathbf{a}}_{qN}^{MCO}\simeq\mathbf{N}\left(\mathbf{a}_{q0},\frac{\hat{\mathbf{\Sigma}}_N}{N}\right).$$

3. Conséquence 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 :

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

ou, en version « compacte » :

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

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 »

Si la variable q_i est omise, ses effets « en moyenne » se retrouvent dans la constante λ et ses effets « hors-moyenne » se retrouvent dans le terme d'erreur u_i. Le modèle initial (ou structurel latent, puisqu'il donne la structure de la relation même si cette structure n'est pas complètement observable donc en partie latente) et donné par :

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

Le modèle utilisé sans l'effet de q_i, ou modèle observable, est donné par :

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

On a donc :

$$y_i = \alpha + \beta n_i + \delta q_i + v_i = \lambda + \beta n_i + u_i$$

d'où :

$$\alpha + \delta q_i + v_i = \lambda + u_i .$$

Puisque $E\left[u_i\right] = 0$, on a :

$$\lambda = \alpha + \delta E\left[q_i\right] \text{ et } u_i = v_i + \delta(q_i - E\left[q_i\right]) .$$

- Donc le paramètre β a toujours la même interprétation. En revanche le terme d'erreur u_i et la constante λ ont des interprétations différentes.
- La constante λ est la somme du salaire de base moyen α et de l'effet de l'agilité intellectuelle moyenne $E\left[q_i\right]$, i.e. $\delta E\left[q_i\right]$.
- Le terme d'erreur u_i est la somme du terme d'erreur v_i et de l'effet de l'agilité intellectuelle hors-moyenne $q_i - E\left[q_i\right]$, i.e. $\delta(q_i - E\left[q_i\right])$.

- *Remarque* : q_i est parfois nommée variable d'hétérogénéité latente.

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

- En utilisant, la forme compacte du modèle observable :

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

il est facile d'écrire l'estimateur des MCO de $\mathbf{a}_0$:

$$\hat{\mathbf{a}}_N^{MCO} = \begin{bmatrix} \hat{\lambda}_N^{MCO} \\ \hat{\beta}_N^{MCO} \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i y_i \,.$$

- Il est toujours possible de calculer cet estimateur, mais cet estimateur ne converge pas toujours vers $\mathbf{a}_0$. Certaines conditions doivent être vérifiées pour cela.

- En utilisant les notations introduites et la formule de l'estimateur des MCO, il est possible de montrer que l'estimateur des MCO de β (notre paramètre d'intérêt) s'écrit sous la forme :

$$\hat{\beta}_N^{MCO} = \frac{\frac{1}{N}\sum_{i=1}^{N}(n_i - \overline{n})(y_i - \overline{y})}{\frac{1}{N}\sum_{i=1}^{N}(n_i - \overline{n})^2},$$

c'est l'estimateur de régression du coefficient de pente dans un modèle linéaire à une variable explicative (et une constante).

Complément

Pour cela il suffit de remarquer que :

$$\mathbf{x}_i\mathbf{x}_i' = \begin{bmatrix} 1 & n_i \\ n_i & n_i^2 \end{bmatrix}$$

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i' = \begin{bmatrix} 1 & \overline{n} \\ \overline{n} & \frac{1}{N}\sum_{i=1}^{N}n_i^2 \end{bmatrix}$$

$$\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right]^{-1} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}n_i^2 - \overline{n}^2}\begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}n_i^2 & -\overline{n} \\ -\overline{n} & 1 \end{bmatrix} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}(n_i - \overline{n})^2}\begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}n_i^2 & -\overline{n} \\ -\overline{n} & 1 \end{bmatrix}$$

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i y_i = \begin{bmatrix} \overline{y} \\ \frac{1}{N}\sum_{i=1}^{N}n_i y_i \end{bmatrix}$$

$$\hat{\beta}_N^{MCO} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}(n_i - \overline{n})^2}\begin{bmatrix} -\overline{n} & 1 \end{bmatrix}\begin{bmatrix} \overline{y} \\ \frac{1}{N}\sum_{i=1}^{N}n_i y_i \end{bmatrix}$$

$$= \frac{\left[-\overline{n}\,\overline{y} + \frac{1}{N}\sum_{i=1}^{N}n_i y_i\right]}{\frac{1}{N}\sum_{i=1}^{N}(n_i - \overline{n})^2} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})(n_i - \overline{n})}{\frac{1}{N}\sum_{i=1}^{N}(n_i - \overline{n})^2}$$

- En remplaçant y_i par le modèle transformé $\lambda + \beta n_i + \left(v_i + \delta(q_i - E\left[q_i\right])\right)$, et après simplifications, il est possible d'écrire $\hat{\beta}_N^{MCO}$ sous la forme :

$$\hat{\beta}_N^{MCO} = \beta + \delta \frac{\frac{1}{N}\sum_{i=1}^{N}(n_i - \overline{n})(q_i - E\left[q_i\right])}{\frac{1}{N}\sum_{i=1}^{N}(n_i - \overline{n})^2} + \frac{\frac{1}{N}\sum_{i=1}^{N}(n_i - \overline{n})v_i}{\frac{1}{N}\sum_{i=1}^{N}(n_i - \overline{n})^2}.$$

- Dès lors, il est possible de montrer que cet estimateur converge en probabilité vers :

$$\hat{\beta}_N^{MCO} \xrightarrow[N\to+\infty]{p} \beta + \delta \frac{Cov\left[n_i, q_i\right]}{V\left[n_i\right]} + \frac{Cov\left[n_i, v_i\right]}{V\left[n_i\right]}.$$

- Puisque q_i et n_i sont exogènes dans le modèle initial on a :

$$Cov\left[n_i, v_i\right] = 0$$

et :

$$\hat{\beta}_N^{MCO} \xrightarrow[N\to+\infty]{p} \beta + \delta \frac{Cov\left[n_i, q_i\right]}{V\left[n_i\right]}.$$

- Cette expression montre que l'estimateur des MCO $\hat{\beta}_N^{MCO}$ n'est un estimateur convergent de β que si :

- $\delta = 0$, *i.e.*, la variable q_i n'est pas pertinente pour expliquer y_i (elle est en fait inutile)

ou/et :

- $Cov\left[n_i, q_i\right] = 0$, *i.e.*, la variable omise et la variable explicative ne sont pas corrélées. Dans ce cas, la convergence de l'estimateur des MCO résulte de l'application du théorème de Frisch-Waugh : dans un modèle linéaire l'omission d'une variable pertinente n'est pas gênante si elle est orthogonale aux autres variables explicatives du modèle.

- Dans tous les autres cas, c'est-à-dire lorsque $\delta \neq 0$ et $Cov\left[n_i, q_i\right] \neq 0$, $\hat{\beta}_N^{MCO}$ ne converge pas vers β.

Complément

- Cas $\delta = 0$:

Le modèle initial se simplifie en :

$$y_i = \alpha + \beta n_i + v_i \quad \text{et} \quad E\left[v_i / n_i\right] = E\left[v_i\right] = 0$$

et l'hétérogénéité de la relation entre y_i et n_i est contenue dans v_i. Les paramètres de ce modèle peuvent être estimés par les MCO.

Dans ce cas, l'introduction de q_i n'a en fait aucun intérêt.

- Cas $Cov\left[n_i, q_i\right] = 0$:

Le modèle initial peut être transformé en le modèle observable :

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

où $\lambda \equiv \alpha + \delta E\left[q_i\right]$. L'hétérogénéité résiduelle de la relation entre y_i et n_i est contenue dans $u_i = v_i + \delta\left(q_i - E\left[q_i\right]\right)$. Ce modèle est un modèle linéaire à variables explicatives exogènes.

Sa seule particularité est que ses paramètres doivent être analysés de manière un peu particulière, la constante l'occurrence. Dans ce cas le paramètre de la constante inclut l'effet moyen de la variable omise $\delta E\left[q_i\right]$. Les paramètres de ce modèle peuvent être estimés par les MCO puisque n_i est exogène dans le modèle observable.

En fait ici, les effets « hors-moyenne » de q_i se trouvent dans u_i mais, puisqu'ils ne sont pas liés à n_i, cela ne pose pas de problème d'endogénéité de n_i.

- Cas $\delta \neq 0$ et $Cov\left[n_i, q_i\right] \neq 0$:

Dans ce cas, q_i est à la fois pertinente dans la relation étudiée et corrélée à la variable explicative utilisée. Il est toujours possible de transformer le modèle comme dans le cas précédent :

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

où : $\lambda \equiv \alpha + \delta E\left[q_i\right]$ et $u_i = v_i + \delta\left(q_i - E\left[q_i\right]\right)$. Mais dans ce cas, u_i et n_i sont corrélés puisque u_i contient les effets de q_i qui sont corrélés à n_i. Aussi, n_i ne peut donc être considérée comme exogène dans le modèle observable.

Dans ce cas, $\hat{\beta}_N^{MCO}$ n'est plus convergent puisque n_i n'est pas exogène par rapport à u_i.