Notes d'économétrie

Auteur :

Pierre Rouillard

Last update : 27 mai 2023

Table des matières

1 Econo2

1.1 Interprétation du paramètre causal à estimer

Selon le modèle considéré il est possible ou non d'avoir une interprétation quantitative directe et/ou qualitative du paramètre causal à estimer β_0.

Définition de l'effet marginal de X_k sur Y : $\frac{\partial E[Y|X=x]}{\partial x_k}$

↪ **Modèle linéaire :**

Comme toujours par la suite on considère l'analyse *toutes choses égales par ailleurs, sur la population considérée...*

Interprétation : la variable d'intérêt est Y et en l'absence de puissance ou d'interactions on peut interpréter quantitativement β_0 sur la variable d'intérêt. Le paramètre d'intérêt est l'effet marginal de X_k sur Y qui vaut bien β_{0k} lorsque X_k apparaît simplement dans le modèle. C'est justement pour cela qu'on peut bien interpréter directement quantitativement les coefficients de β_0 !

↪ **Modèle binaire :**

$E[Y|X] = P(Y = 1|X) = F(X'\beta_0)$

$E[Y|X] = F(X'\beta_0) \iff Y = \mathbb{1}(Y^* \geq s) \ : \ Y^* = X'\beta_0 + \varepsilon \quad \varepsilon \perp\!\!\!\perp X$

Interprétation quantitative directe : la variable d'intérêt est Y et Y^* n'est qu'une variable latente qui n'a pas forcément de sens quantitatif précis. Le paramètre d'intérêt est l'effet marginal de X_k sur la variable d'intérêt, ici Y. Les coefficients de β_0 concernant Y^* on ne peut donc pas directement interpréter quantitativement ces derniers sur la variable d'intérêt Y. De plus, l'effet marginal de X_k sur Y est différent de β_{0k} : c'est pour cela qu'on ne peut avoir d'interprétation quantitative des coefficients de β_0 !

Interprétation qualitative : en revanche le signe de l'effet marginal de X_k sur Y, i.e. effet positif ou négatif sur $P(Y = 1|X)$, est donné par le signe de β_{0k}.

On peut en revanche comparer quantitativement le ratio des effets marginaux des variables i et j qui vaut $\widehat{\beta_i}/\widehat{\beta_j}$. *L'effet sur la proba d'être ... de la variable i est <quantitativement> ... que l'effet de la variable j* $\iff$ *regarder le rapport* $\widehat{\beta_i}/\widehat{\beta_j}$.

↪ **Modèle de censure / Tobit1 :**

1 seul mécanisme détermine la valeur de Y et si on observe la variable d'intérêt ou non. Deux cas sont à distinguer :

⇒ **Données censurées** : la variable d'intérêt est Y^* qui peut ne pas être observée au dessous d'un seuil causant un problème de censure. Le paramètre d'intérêt est l'effet marginal de X_k sur la variable d'intérêt Y^*, qui vaut bien β_{0k} lorsque X_k apparaît simplement dans

le modèle linéaire de Y^*. Ainsi, la variable Y^* ayant un sens quantitatif et malgré la censure liée aux problèmes d'observation on peut bien interpréter quantitativement β_0 sur la variable d'intérêt.

$\Rightarrow$ **Solution en coin** : la variable d'intérêt est bien Y alors que la variable Y^* est une variable latente potentiellement dépourvue de sens quantitatif. Typiquement un pb d'optimisation du consommateur où Y^* mesure l'utilité optimale (en nombre de biens) de consommation d'un bien donné : donc potentiellement négatif. Et Y représente le nombre d'unités effectivement consommées. Les coefficients de β_0 concernant Y^* qui n'as pas de sens quantitatif précis : on ne peut pas interpréter quantitativement les coefficients de β_0 sur la variable d'intérêt Y. Les paramètres d'intérêt sont les effets marginaux : le total $\frac{\partial E[Y|X=x]}{\partial x_k}$ (marge extensive et intensive) et $\frac{\partial E[Y|Y>0,X=x]}{\partial x_k}$ (marge intensive seulement). Ces paramètres sont tous les deux différents de β_{0k} ce qui explique le manque d'interprétation quantitative des coefficients de β_0.

↪ **Modèle de sélection / Tobit2 :**

Ici on a bien deux processus différents : un qui détermine Y^* **et un autre** qui détermine si on observe cette valeur ou non i.e. modèle sur D.

Interprétation quantitative directe : il y a un problème d'observation des données, on observe $Y = D.Y^*$ mais la variable d'intérêt est bien Y^* (variable potentielle qui existe pour tous les *individus*). Par conséquent Y^* suivant un modèle linéaire, les paramètres d'intérêts sont les effets marginaux des variables explicatives sur la variable d'intérêt Y^* et les coefficients de β_0 sont toujours interprétables quantitativement.

1.2 Représentations linéaires & OLS

On note $Y \in \mathbf{R}$ la variable d'intérêt/dépendante et $X \in \mathbf{R^K}$ le vecteur des variables explicatives.

↪ **Représentation non causale - Projection linéaire :**

Sous conditions de moments[1], on a toujours par construction/définition de la représentation linéaire théorique (=projection linéaire orthogonale) orthogonalité des *résidus* de cette représentation non causale avec les régresseurs = pas une hypothèse mais une conséquence.

$Y = X'.\widetilde{\beta} + \widetilde{\varepsilon}$, $E[X\widetilde{\varepsilon}] = 0$ toujours définissable sous conditions de moments.

- $\widehat{\beta}_{OLS}$ estime toujours $\widetilde{\beta}$: $\widehat{\beta}_{OLS} \underset{n\to+\infty}{\longrightarrow} \widetilde{\beta}$
- $X'.\widetilde{\beta}$ meilleure prédiction linéaire de Y par X : $\widetilde{\beta}$ solution MSE.

$$\widehat{\beta}_{OLS} \in \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - X_i'.\beta)^2 \quad \longleftrightarrow \quad \widetilde{\beta} \in \underset{\beta}{\operatorname{argmin}}\, E[(Y - X'.\beta)^2]$$

↪ **Représentation causale :**

La représentation causale fait intervenir le paramètre causal β_0 qu'on cherche à estimer : dans cette représentation le *terme d'erreur* n'est pas automatiquement orthogonal au régresseur.

$Y = X'.\beta_0 + \varepsilon$, $E[X\varepsilon] \overset{?}{=} 0$

- β_0 paramètre causal à estimer.
- ε résidu : agrège les facteurs inobservés qui affectent Y.

Le terme d'erreur ε capte l'hétérogénéité inobservée , i.e capte les déterminants inobservés qui affectent la variable d'intérêt Y : deux individus avec les mêmes variables explicatives auront néanmoins la plupart du temps des variables expliquées différentes.

Avoir orthogonalité ($\to$ indépendance) entre régresseurs et terme d'erreur est une hypothèse ! C'est **l'hypothèse d'exogénéité**.

↪ **Lien :**

Sans l'hypothèse d'exogénéité pour la représentation causale, les deux représentations diffèrent et l'estimateur OLS ne permet pas d'identifier le paramètre causal d'intérêt.

En revanche avec hypothèse d'exogénéité les deux représentations coïncident et $\widetilde{\beta} = \beta_0$: $\widehat{\beta}_{OLS}$ qui estime toujours $\widetilde{\beta}$ est donc un estimateur consistant de β_0.

1. $E[Y^2] < +\infty$, $E[\|X\|^2] < +\infty$ et $E[XX']$ inversible = de rang plein. En particulier : **any level of correlation between covariates except perfect colinearity** : composantes de X linéairement indépendantes mais *n'exclut pas* qu'elles soients corrélées. Si le modèle a une constante et une variable catégorielle il faut exclure une des modalités.

En dehors des expériences contrôlées les variables explicatives peuvent parfois être corrélées aux facteurs inobersables et pb d'endogénéité $E[X\varepsilon] \neq 0$.

↪ **Mémo OLS :**

↻ *link-ols*

Représentation causale

Estimateur OLS - MCO

$Y = X'.\beta_0 + \varepsilon$, avec $E[\varepsilon] = 0$

- Conditions de moments
- $E[X\varepsilon] = 0$ hypothèse d'exogénéité

$\Rightarrow \beta_0$ est identifiable $\beta_0 = E[XX']^{-1}E[XY]$

$\Rightarrow$ L'estimateur $\widehat{\beta}_{OLS}$ est consistant, sans biais & asymptotiquement normal

On a :

$$\widehat{\beta}_{OLS} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)$$

$$\sqrt{n}(\widehat{\beta}_{OLS} - \beta_0) \underset{n\to+\infty}{\longrightarrow} \mathcal{N}(O, E[XX']^{-1}E[\varepsilon^2 XX']E[XX']^{-1})$$

Homoscédasticité : hypothèse sur la variance des résidus. L'erreur standard des MCO diffèrent en fonction de l'hypothèse !

- Faible $E[\varepsilon^2 XX'] = E[\varepsilon^2]E[XX']$
- Forte $E[\varepsilon^2|X] = \sigma^2 \leftrightarrow V[\varepsilon|X] = \sigma^2$

L'hypothèse d'homoscédasticité forte requiert que la variance des termes d'erreur soit la même pour chaque observation. L'estimateur $\widehat{V}$ de la variance asymptotique de $\widehat{\beta}_{OLS}$ est robuste à l'hétéroscédasticité. L'estimateur standard $\widetilde{V}$ n'est convergent que sous l'hypothèse d'homoscédasticité.

$$\underline{\textit{Robust}} \quad \widehat{V} = \left(\frac{1}{n-k}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n-k}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 X_i X_i'\right)\left(\frac{1}{n-k}\sum_{i=1}^{n} X_i X_i'\right)^{-1}$$

$$\underline{\textit{Standard}} \quad \widetilde{V} = \left(\frac{1}{n-k}\sum_{i=1}^{n} \hat{\varepsilon}_i^2\right)\left(\frac{1}{n-k}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \quad \text{sous hypothèse d'homoscédasticité}$$

1.3 Modèle I.V - estimateur 2SLS/2MC

↻ *link-2sls*

On suppose $Y = X'.\beta_0 + \varepsilon$ avec problème d'**endogénéité** $E[X\varepsilon] \neq 0$

Variables instrumentales : $X \in \mathbf{R^K}$, $Z \in \mathbf{R^L}$, avec $L \geqslant K$

1. Exogénéité : $E[Z\varepsilon] = 0$
2. Condition de rang $E[ZX']$ de rang $K \leftrightarrow$ Pertinence $Cov(Z, X^{(i)}) \neq 0 \quad \forall i$

Condition de rang donne qu'il exite Γ tq $\Gamma.E[ZX']$ inversible. *Condition de rang* testable avec first step (significativité d'un des coeff. des vrais instruments) et *exogénéité* pas testable.
Si $L = K$ alors β_0 est *juste identifié*, si $L > K$ alors β_0 est *suridentifié*.
Z inclut toutes les variables exogènes. En pratique on ne régresse et remplace par le projection linéaire sur Z estimée que les variables endogènes.

Estimateur 2MC - 2SLS

1. Projection linéaire de X sur Z $\mapsto X^* = \Gamma Z$
   où $\Gamma = E[XZ']E[ZZ']^{-1} = (\beta^{OLS'}_{(1)/Z}, ..., \beta^{OLS'}_{(K-1)/Z})^T = "\beta^{OLS}_{X/Z}"$
2. Reg lin de Y sur $X^* \mapsto$ OLS estimé est $\widehat{\beta}_{2SLS}$

. $Y = X'.\beta_0 + \varepsilon$

. $\beta_0 = E[\Gamma ZX']^{-1}E[\Gamma ZY] = E[X^*X']^{-1}E[X^*Y]$

or $\langle z, x - p_{\mathbf{Z}}(x)\rangle = 0 \quad \forall z \in \mathbf{Z} \quad \Rightarrow \quad z = X^* \in Vect(Z) \quad E[X^*X^T] = E[X^*X^{*T}]$

. $\beta_0 = E[X^*X^{*'}]^{-1}E[X^*Y]$

Estimateur doubles moindres carrés : $\hat{\beta}_{2SLS} \underset{n\to+\infty}{\longrightarrow} \beta_0$

$$\hat{\beta}_{2SLS} = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{X}_i\hat{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{X}_iY_i\right) \quad \text{où} \quad \hat{X}_i = \hat{\Gamma}Z_i \quad \hat{\Gamma} = (\hat{\beta}^{OLS'}_{(1)/Z}, ..., \hat{\beta}^{OLS'}_{(K-1)/Z})^T$$

$\Rightarrow$ L'estimateur $\hat{\beta}_{2SLS}$ est convergent & asymptotiquement normal **mais** pas necéssairement sans biais.

$$\sqrt{n}(\widehat{\beta}_{2SLS} - \beta_0) \underset{n\to+\infty}{\longrightarrow} \mathcal{N}(O, \mathrm{VA}(\widehat{\beta}_{2SLS}) = \underline{E[X^*X^{*'}]^{-1}E[\varepsilon^2X^*X^{*'}]E[X^*X^{*'}]^{-1}})$$

$\hookrightarrow$ ε estimé par $\widehat{\varepsilon}_i = Y_i - \boldsymbol{X_i'}.\widehat{\beta}_{2SLS}$ et $\widehat{\mathrm{VA}}(\widehat{\beta}_{2SLS})$ estimateur convergent robuste (car ne repose pas sur des hypothèses d'homoscédasticité) de la variance asymptotique. Ce n'est pas l'estimateur qu'on obtient si l'on fait une régression de Y sur $\widehat{X}$ car $\widehat{\varepsilon}_i \neq Y_i - \widehat{X}_i'.\widehat{\beta}_{2SLS}$

1.4 Méthode des moments généralisés

↻ *link-gmm*

$(U_i = (Y_i, X_i, Z_i))$ l'ensemble des données sur i

On veut estimer $\theta_0 \in \mathbf{R^K}$ en utilisant : $E[g(U, \theta_0)] = 0 \quad g \mapsto \mathbf{R}^L,\ L \geqslant K$

Estimtateur GMM pas unique dans le cas $L > K$ car dépend de la matrice de pondération $\widehat{W}_n$ choisie.

Hypothèse d'identification ⇒ GMM convergent et asymptotiquement normal.

Choix théorique $W_0 = H^{-1} = V(g(U, \theta_0))^{-1} = E[g(U, \theta_0).g(U, \theta_0)^T]^{-1}$ est optimal : inconnu car fonction de $\theta_0 \mapsto W_n = \widehat{W}_0$

GMM est optimal ↦ on sous-entend qu'on choisit $W_n \underset{n\to+\infty}{\longrightarrow} W_0 = H^{-1}$ *matrice de pondération optimale.* Asymptotiquement efficace/optimal

Cas exogène : $g(U, \beta) = X(Y - X'.\beta)\ , \quad E[g(U, \beta_0)] = 0 \quad L = K$

Tous les choix de $\widehat{W}_n$ *matrice de pondération* (symétrique, définie potentiellement aléatoire) conduisent au même estimateur = MCO

Cas instrumental : $g(U, \beta) = Z(Y - X'.\beta)\ , \quad E[g(U, \beta_0)] = 0 \quad Z \in \mathbf{R}^L \quad \mathbf{L>K}$

Différents $W_n \longleftrightarrow$ différents estimateurs.

- **Cas Homoscédastique :** l'estimateur GMM coïncide avec l'estimateur des 2MC (2SLS) ⇒ 2SLS est optimal dans le cas homoscédastique
- Sinon présence d'**hétéroscédastique** dans les résidus : GMM donne un nouvel estimateur ≠2SLS qui lui est bien optimal.

2 XXX

2.1 xxx