Econométrie des données de panel

Guillaume Horny*

*Banque de France

Master 2 MASERATI

Conclusion

Plan

Plan

Variable aléatoire ou paramètre ?

Modèle de base :

$$y_{it} = x_{it}^{'}\beta + \alpha_i + \epsilon_{it},$$

On a vu que les α_i peuvent être considérés comme :

- des paramètres (modèle à effets fixes, chapitre 1)
- ou comme les réalisation d'une variable aléatoire (modèle à effets aléatoires, chapitre 2).

Ce choix de modélisation conduit à faire des hypothèses différentes, aboutissant à des estimateurs différents.

On va voir ici que tous les estimateurs vus précédemment appartiennent à une même famille : celle des **estimateurs de classe** λ.

Plan

Estimateurs de classe λ

Un estimateur de classe λ est tel que :

$$\widehat{\beta}(\lambda) = \left[X'(W + \lambda B)X\right]^{-1} X'(W + \lambda B)Y,$$

où :

- $W = I_{NT} - X_1(X_1'X_1)^{-1}X_1'$. Cette matrice était notée M_{X_1} dans le chapitre 1, avec X_1 les indicatrices d'individus,
- λ est un scalaire,
- $B = X_1(X_1'X_1)^{-1}X_1'$.

Ecriture de B (1/1)

En effet, lorsque X_1 est la matrice des indicatrices d'individus :

$$X_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Ecriture de B (2/2)

$$(X_1'X_1)^{-1} = \begin{pmatrix} T & 0 & \cdots & 0 \\ 0 & T & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & T \end{pmatrix}^{-1} = \frac{1}{T} I_N.$$

$$X_1(X_1'X_1)^{-1}X_1' = \frac{1}{T} X_1 X_1' = \frac{1}{T} \begin{pmatrix} 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \end{pmatrix}$$

Ecriture de B (3/3)

On a :

$$BY = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} y_{1t} \\ \vdots \\ \frac{1}{T}\sum_{t=1}^{T} y_{1t} \\ \vdots \\ \vdots \\ \frac{1}{T}\sum_{t=1}^{T} y_{Nt} \\ \vdots \\ \frac{1}{T}\sum_{t=1}^{T} y_{Nt} \end{pmatrix}$$

$\Rightarrow$ il s'agit de la transformation *between* de Y. De même, on peut montrer que BX_2 est la transformation *between* de X_2.

Régression *between* (1/3)

L'estimateur *between* est l'estimateur OLS appliqué au modèle :

$$BY = BX\beta + B\epsilon.$$

Son écriture matricielle est :

$$\begin{aligned}\widehat{\beta}_{OLS} &= \left((BX)'BX\right)^{-1}(BX)'BY \\ &= \left(X'B'BX\right)^{-1}X'B'BY \\ &= \left(X'BX\right)^{-1}X'BY.\end{aligned}$$

car B est carrée, symétrique et idempotente ($B = B^2$).

Régression *between* (2/3)

On a vu dans le chapitre 2 :

" L'estimateur between est l'estimateur OLS de la régression de $y_{i.}$ sur une constante et les moyennes intertemporelles $x_{i.}$ (N observations !)."

Or, BY est de dimension $(NTx1)$, et non pas $(Nx1)$? ! ?

On peut montrer que :

- la régression de BY sur BX et une constante, où BY est de dimension $(NTx1)$,
- celle du vecteur des $y_{i.}$ (de dimension $(Nx1)$), sur $x_{i.}$ et une constante,

conduisent au même estimateur $\widehat{\beta}$.

Régression *between* (3/3)

Par contre, comme le vecteur des résidus n'est pas de même dimension, l'estimateur de $\text{Var}(\beta) = \text{E}(\epsilon\epsilon')$ ne sera pas identique. Faire la régression sur un vecteur $(Nx1)$ est un moyen simple de corriger pour la corrélation des erreurs (répéter T fois une observation n'apporte pas d'information !).

Ecriture de W

Ainsi :

$$WY = \left(I_{NT} - \frac{1}{T} X_1 X_1^{'} \right) Y = \begin{pmatrix} y_{11} - \frac{1}{T} \sum_{t=1}^{T} y_{1t} \\ \vdots \\ y_{1T} - \frac{1}{T} \sum_{t=1}^{T} y_{1t} \\ \vdots \\ y_{N1} - \frac{1}{T} \sum_{t=1}^{T} y_{Nt} \\ \vdots \\ y_{NT} - \frac{1}{T} \sum_{t=1}^{T} y_{Nt} \end{pmatrix}.$$

WY est donc l'écriture matricielle de la transformée *within* de Y. De même, WX_2 est la transformée *within* de X_2.

Retour sur l'estimateur de classe λ

Estimateur de classe λ :

$$\widehat{\beta}(\lambda) = \left[X'(W + \lambda B)X\right]^{-1} X'(W + \lambda B)Y.$$

On a :

- $\widehat{\beta}(0) = \widehat{\beta}_{Within}$,
- $\widehat{\beta}(1) = \widehat{\beta}_{OLS}$, car $W = I_{NT} - B$,
- $\widehat{\beta}(\infty) = \widehat{\beta}_{Between}$,
- $\widehat{\beta}\left(\frac{\widehat{\sigma}_w^2}{\widehat{\sigma}_w^2 + T\widehat{\sigma}_\alpha^2}\right) = \widehat{\beta}_{FGLS}$,
- $\widehat{\beta}\left(\frac{\sigma_w^2}{\sigma_w^2 + T\sigma_\alpha^2}\right) = \widehat{\beta}_{GLS}$.

Estimateurs de classe λ : le cas *between*

Lorsque $\lambda \to \infty$, W devient négligeable par rapport à B. D'où :

$$\begin{aligned}\widehat{\beta}(\lambda) &= \left[X'(\lambda B)X\right]^{-1} X'(\lambda B)Y \\ &= \left[X'BX\right]^{-1} X'BY.\end{aligned}$$

L'écriture ci-dessus n'est pas ce qu'on a fait de plus propre mathématiquement. L'idée est juste que la composante *within* (intra-individuelle) est écrasée par la composante *between* (inter-individuelle).

Estimateurs de classe λ : les cas FGLS et GLS

- **Estimateur FGLS**
  On a vu dans le chapitre 2 que l'estimateur FGLS est équivalent à l'estimateur OLS du modèle transformé :

$$y_{it} - \widehat{\lambda} y_{i.} = (1 - \widehat{\lambda})\beta_0 + (x_{it} - \widehat{\lambda} x_{i.})' \beta + v_{it},$$

  où $\widehat{\lambda}$ est un estimateur convergent de $\lambda = 1 - \sigma_w^2 / (T\sigma_\alpha^2 + \sigma_w^2)$.
- **Estimateur GLS**
  L'estimateur GLS est obtenu lorsqu'on dispose de la vraie valeur de λ.

Plan

Principe général

On a vu que :

- sous l'hypothèse de modèle à **effet fixe**, c'est-à-dire de corrélation potentielle entre les caractéristiques observables et inobservables, on dispose d'estimateurs basés sur des transformations des données qui sont convergents mais pas efficaces
- sous l'hypothèse de modèle à **effet aléatoire**, c'est-à-dire sous l'hypothèse d'absence de corrélation entre les caractéristiques observables et inobservables, on dispose d'un estimateur FGLS convergent et efficace.

On cherche maintenant des estimateurs qui, lorsque les caractéristiques inobservables sont corrélées avec les observables, resteront toujours efficaces.

Le problème

On repart du modèle à effets aléatoires :

$$y_{it} = x_{it}^{'}\beta + \alpha_i + w_{it}.$$

On suppose cette fois-ci que les α_i sont corrélés avec les variables explicatives :

$$\mathrm{E}(\alpha_i | x_{i1}, \ldots, x_{iT}) \neq \mathrm{E}(\alpha_i).$$

Attention, il ne s'agit en aucune manière d'une conséquence des espérances itérées ! (cf Annexe)

Le problème (1/2)

Les autres hypothèses sont valides, à quelques ajustements près :

1. H1b : $\mathrm{E}(w_{it}|x_{i1},\ldots,x_{iT}) = 0, t = 1,\ldots,T.$
2. H2b : $\mathrm{E}(\alpha_i) = 0$
3. H3 :
   - $\mathrm{E}(\alpha_i^2|x_{i1},\ldots,x_{iT}) = \sigma_\alpha^2,$
   - $\mathrm{E}\left[(w_{i1},\ldots,w_{iT})(w_{i1},\ldots,w_{iT})'|x_{i1},\ldots,x_{iT},\alpha_i\right] = \sigma_w^2 I_T.$

On retrouve donc toujours la structure si spécifique de la matrice de variance (équicorrélation).

Le problème (2/2)

La spécificité du modèle à erreurs corrélées est que :

$$\begin{aligned}
\mathrm{E}(y_{it}|x_{i1},\ldots,x_{iT}) &= x_{it}^{'}\beta + \mathrm{E}(\epsilon_{it}|x_{i1},\ldots,x_{iT}) \\
&= x_{it}^{'}\beta + \mathrm{E}(\alpha_i|x_{i1},\ldots,x_{iT}) + \mathrm{E}(w_{it}|x_{i1},\ldots,x_{iT}) \\
&= x_{it}^{'}\beta + \mathrm{E}(\alpha_i|x_{i1},\ldots,x_{iT}) \\
&\neq x_{it}^{'}\beta
\end{aligned}$$

Les caractéristiques inobservables affectent ici l'espérance conditionnelle de y_{it} ainsi que sa variance (cette dernière par H3). Les inobservables introduisent à la fois des écarts dans les moyennes individuelles et une surdispersion.

Les conséquences du problème (1/4)

$$\widehat{\beta}(\lambda) = \left[X'(W+\lambda B)X\right]^{-1} X'(W+\lambda B)Y,$$

D'où :

$$\begin{aligned}
\mathrm{E}[\widehat{\beta}(\lambda)] =& \mathrm{E}\left[\left[X'WX+\lambda X'BX\right]^{-1}(X'W+\lambda X'B)(X\beta+\epsilon)\right] \\
=& \mathrm{E}\left[\left[X'WX+\lambda X'BX\right]^{-1}(X'W+\lambda X'B)X\beta\right] \\
&+ \mathrm{E}\left[\left[X'WX+\lambda X'BX\right]^{-1}(X'W+\lambda X'B)\epsilon\right] \\
=& \beta + \mathrm{E}\left[\left[X'WX+\lambda X'BX\right]^{-1}(X'W+\lambda X'B)\epsilon\right].
\end{aligned}$$

Comme ϵ est corrélé avec X, on ne peut pas écrire $\mathrm{E}[f(X)\epsilon] = \mathrm{E}[f(X)]\mathrm{E}[\epsilon]$.

Les conséquences du problème (2/4)

Poursuivons les calculs :

$$\mathrm{E}[[X'WX + \lambda X'BX]^{-1}(X'W+\lambda X'B)\epsilon] =$$
$$\mathrm{E}\left[[X'WX + \lambda X'BX]^{-1}X'W(\alpha + w)\right]$$
$$+ \mathrm{E}\left[[X'WX + \lambda X'BX]^{-1}\lambda X'B(\alpha + w)\right].$$

W est l'expression matricielle de la transformation *within*. Ainsi, $W\alpha = 0$. De plus, $\mathrm{E}(w) = 0$ et $\mathrm{E}(w|X) = \mathrm{E}(w)$. D'où la disparition du premier terme :

$$\mathrm{E}[[X'WX + \lambda X'BX]^{-1}(X'W + \lambda X'B)\epsilon]$$
$$=\mathrm{E}\left[[X'WX + \lambda X'BX]^{-1}\lambda X'B(\alpha + w)\right]$$
$$=\mathrm{E}\left[[X'WX + \lambda X'BX]^{-1}\lambda X'B\alpha\right]$$
$$+ \mathrm{E}\left[[X'WX + \lambda X'BX]^{-1}\lambda X'Bw\right].$$

Les conséquences du problème (3/4)

Comme w est supposé sans corrélation avec les X :

$$\mathrm{E}\left[[X'WX + \lambda X'BX]^{-1}\lambda X'Bw\right] = 0.$$

Ainsi :

$$\mathrm{E}\left[[X'WX + \lambda X'BX]^{-1}(X'W + \lambda X'B)\epsilon\right] = \mathrm{E}\left[[X'WX + \lambda X'BX]^{-1}\lambda X'B\alpha\right]$$

Au final, on a :

$$\mathrm{E}[\widehat{\beta}(\lambda)] = \beta + \mathrm{E}\left[[X'WX + \lambda X'B]^{-1}\lambda X'B\alpha\right].$$

Pour qu'un estimateur de la famille de classe λ soit sans biais, il faut que le dernier terme soit nul, ce qui arrive lorsque $\lambda = 0$.

Les conséquences du problème (4/4)

De tous les estimateurs de classe λ (estimateur du modèle empilé, *within*, FGLS...), seul l'estimateur *within* est sans biais et convergent lorsque les effets individuels α_i sont corrélés avec les régresseurs.

- Intuitivement, en éliminant la source de la corrélation, c'est-à-dire les α_i, on a éliminé la corrélation entre l'erreur composée et les variables explicatives.
- A noter que la convergence est par ailleurs également vérifiée pour l'estimateur du modèle en **différence première** (cf chapitre 1), qui n'appartient par à la famille des estimateurs de classe λ.

Convergence de $\widehat{\beta}$

	Modèle supposé	
	Effets fixes	Effets aléatoires
Estimateur de β	$\text{corr}(\alpha, X_k) \neq 0$	$\text{corr}(\alpha, X_k) = 0$
Empilé	Non-convergent	Convergent
Within	Convergent	Convergent
Différence première	Convergent	Convergent
Between	Non-convergent	Convergent
Effet aléatoire	Non-convergent	Convergent

Plan

Le test d'Hausman

L'objectif est de tester une hypothèse d'absence de corrélation. Selon son acceptation ou son rejet, on va privilégier certains estimateurs plutôt que d'autres.

Jerry Haussman (1946-), actuellement professeur d'Economie au MIT. A étudié le secteur des télécommunication et les questions de concurrence, de régulation et de taxation, entre autres. Sa contribution la plus connue est le test qui porte son nom, publié en 1978.

Le test d'Hausman : l'idée générale

Nous sommes dans un cas de figure où il existe plusieurs manières d'estimer un modèle. Si une des hypothèses du modèle n'est pas vérifiée, certains estimateurs seront convergents tandis que d'autres ne le seront plus. On va mesurer l'écart qu'il y a entre les deux estimations. Deux possibilités :

- les deux estimations sont **similaires**, l'écart est proche de 0, les données ne semblent pas contredire l'hypothèse testée,
- les deux estimations sont **différentes**, l'écart est significativement différent de 0, l'hypothèse testée ne semble pas satisfaite par les données.

Application à des données de panel

Sous l'hypothèse de corrélation entre les caractéristiques inobservables et les variables explicatives, l'estimateur FGLS n'est plus convergent alors que l'estimateur *within* reste convergent. On peut donc comparer les deux estimations.

- Si elles sont voisines, on peut accepter l'hypothèse d'exogénéité des X par rapport à α. On n'a donc pas de problème d'endogénéité des X par rapport au terme d'erreur composé.
- A l'inverse, si les deux estimations son très différentes, au moins l'une des estimations n'est pas convergente et on rejette l'hypothèse d'exogénéité des X par rapport à α.

Statistique de test

Dans le cas des panels, la statistique de test originelle est :

$$S_H = (\widehat{\beta}_{within} - \widehat{\beta}_{FGLS}) \left[\text{Var}(\widehat{\beta}_{within}) - \text{Var}(\widehat{\beta}_{FGLS}) \right]^{-1} (\widehat{\beta}_{within} - \widehat{\beta}_{FGLS}).$$

La soustraction des variances conduit parfois à des problèmes numériques dans certaines applications. Une statistique alternative a été proposée par Hausman et Taylor en 1981 :

$$S_{HT} = (\widehat{\beta}_{between} - \widehat{\beta}_{within}) \left[\text{Var}(\widehat{\beta}_{between}) + \text{Var}(\widehat{\beta}_{within}) \right]^{-1} (\widehat{\beta}_{between} - \widehat{\beta}_{within}).$$

Elle suit un χ^2 à autant de degrés de libertés qu'il y a de variables explicatives dans le modèle après transformation *within*.

Exemple de test d'Hausman (R)

```
R> data("Gasoline", package = "plm")
R> wi <- plm(lgaspcar ~ lincomep + lrpmg + lcarpcap,\\
  data = Gasoline, model = "within")
R> re <- plm(lgaspcar ~ lincomep + lrpmg + lcarpcap,\\
  data = Gasoline, model = "random")
R> phtest(wi, re)

        Hausman Test

data:  lgaspcar ~ lincomep + lrpmg + lcarpcap
chisq = 302.8037, df = 3, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

Exemple de test d'Hausman (Stata)

```
St> xtreg lwage exp exp2 wks ed, fe
St> estimates store FE
St> xtreg lwage exp exp2 wks ed, re
St> estimates store RE
St> hausman FE RE, sigmamore
<SNIP>
 Test:  Ho:  difference in coefficients not systematic
        chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                  =  1513.02
              Prob>chi2  =  0.0000
```

Plan

Les solutions au problème

La présence de corrélation entre les α_i et les x_{it} implique une corrélation entre l'erreur composée et les variables explicatives. De ce point de vue, il s'agit d'un problème classique d'endogénéité. Différentes approches ont été proposées :

- "corriger" le modèle pour évacuer cette corrélation
- avoir recours à des techniques de variables instrumentales (cf Annexe)

Plan

Petite histoire des pratiques

- Historiquement, les modèles à effets fixes ont été les premiers à voir été estimés. Les méthodes pour les modèles à effets aléatoires sont venues ensuite.
- Les modèles à effet aléatoire ont d'abord suscité un fort engouement, avant que les travaux empiriques ne reviennent aux modèle à effet fixe.
- Bien que moins précis, les estimateurs du modèle à efft fixe restent convergent en présence de corrélation entre les caractéristiques inobservables et les variables explicatives. De plus, les gains de précision associés aux estimateurs FGLS sont souvent considérés comme minimes. L'arbitrage coût-bénéfice penche donc, dans la pratique de l'économétrie des panels, souvent en faveur des méthodes les plus simples.

La pratique de l'économétrie des panels aujourd'hui

La littérature empirique contemporaine accorde beaucoup d'importance à l'évaluation des **effets causaux** d'une variable X_k sur la variable dépendante Y. Un grand nombre de projets de recherche utilisent des estimateurs *within*, et portent un soin tout particulier aux questions :

- d'endogénéité de X_k (utilisation de variables instrumentales),
- à la mesure des écarts-types des coefficients (estimations successives avec multiples corrections pour différentes structures d'hétéroscédasticité et d'autocorrélation des erreurs).

Quelle utilité pour les FGLS ?

Dans le cadre d'une utilisation non-académique, l'arbitrage est moins clair. *In fine*, tout va dépendre du domaine d'application, de la question posée, ainsi que de la richesse des données disponibles.

Très grossièrement, on peut dire :

- plus le phénomène étudié est maîtrisé (ex : résultat d'un protocole médical, d'un processus mécanique), plus on dispose d'un modèle détaillé, plus il y a de chance qu'un modèle à effet aléatoire soit convaincant.
- A l'inverse, plus les comportements sous-jacents nous échappent ou plus les variables explicatives potentielles sont rares, plus les inobservables vont jouer un rôle important et plus il est probable qu'elles soient corrélées avec les (quelques) variables explicatives.

Plan

Organisation d'une démarche typique (1/5)

- Si vous avez la possibilité de choisir les questions sur lesquelles vous allez devoir travailler, **réfléchissez bien au type de données qu'il vous faudrait pour pouvoir répondre à votre question**.
  On a besoin de suivre des individus pour étudier l'évolution de leurs comportements, d'observer les offreurs et les demandeurs pour étudier l'équilibre d'un marché, le temps que passent les individus dans un état pour étudier la probabilité qu'ils en sortent...
- **Connaissez vos données avant de faire quoi que ce soit !**
  Comment ont-elles été collectées ? déclaration obligatoire ou volontaire ? qui les a déclarées ? combien de temps après l'événement qui vous intéresse ? à quel point sont-elles fiables ? quelle est la période couverte ? la fréquence d'observation ? si vous avez un panel, la variance est-elle plutôt au niveau intra- ou inter- individuel ?

Organisation d'une démarche typique (2/5)

- **Si vous avez des données de panel, nettoyez les scrupuleusement**. Chercher les observations aberrantes, la forme de l'attrition... **Eviter de cylindrer !** Vous perdrez des observations et restreignez l'échantillon aux individus qui répondent tout le temps (les plus vieux, qui restent tout le temps dans l'état étudié...), alors qu'ils sont généalement loin d'être représentatifs.
- Avant d'écrire votre modèle, revenez à la question de départ. Quel paramètre du modèle permet d'y répondre ? Quelle prévision ? Quelle distribution ? Qui sont les individus concernés ?

Organisation d'une démarche typique (3/5)

- Quelle est forme de **l'hétérogénéité inobservée** ? Entre individus ? entre groupes d'individus ? Est-ce qu'elle affecte le niveau de Y ? sa variance ? les deux ? Sa mesure a-t'elle une importance ?
- Dans le cas général, restez prudents et partez du principe que les **inobservables sont corrélés avec les observables**. Utilisez les estimateurs OLS, *within*, du modèle en différence, FGLS.
- Les résultats sont-ils crédibles ? Comparer les écarts entre les estimations et essayer de les expliquer. D'où vient le problème expliquant les écarts ? Faites des tests d'Hausman, comparer les R^2, regardez les profils des erreurs...

Organisation d'une démarche typique (4/5)

- Les **tests d'effets individuels** conduisent-ils à les garder ?
- Soyez attentifs aux problèmes **d'hétéroscédasticité et d'autocorrélation** des erreurs
- les résultats sont-ils **robustes** à l'introduction de nouvelles variables ? Au changement du mode de calcul de la variable dépendante ou d'une des explicatives ?

Organisation d'une démarche typique (5/5)

- C'est à ce moment seulement que vous pouvez privilégier un estimateur plutôt qu'un autre ! Effectuer votre classement sur la base de **tous** les critères précédents. Quand vos résultats sont obtenus, n'oubliez pas que vous traiter un problème économique qui doit avoir une réponse en termes **économiques**, et pas seulement en terme de coefficients estimés ou de t de Student.
- Si votre problème est de type **causal**, réfléchissez attentivement aux hypothèses d'exogénéité et à leur bien-fondé. Si elles ne tiennent pas pour une variable, comment traiter ce problème **d'endogénéité** ? Est-ce qu'on disposerait d'instruments ? Seraient-ils valides ? seraient-ils faibles ? Auraient-ils un sens économique ? Si oui, tester-les.

Bonne chance !

Annexe

Rappel : loi des espérances itérées (1/2)

La loi des espérances itérées nous dit que $\mathrm{E}_X[\mathrm{E}(Y|X)] = \mathrm{E}(Y)$. Si $\mathrm{E}(Y|X) = 0$, alors $\mathrm{E}_X[\mathrm{E}(Y|X)] = \mathrm{E}_X[0] = 0$. D'où $\mathrm{E}(Y|X) = \mathrm{E}(Y)$. Dire que si $\mathrm{E}(Y) = 0$, alors $\mathrm{E}(Y|X) = 0$, est faux !

Intuitivement : Supposons que X soit une variable aléatoire discrète prenant les valeurs c_1 et c_2 avec probabilités p_1 et p_2. Alors

$$\begin{aligned}\mathrm{E}_X[\mathrm{E}(Y|X)] &= p_1\mathrm{E}(Y|X = c_1) + p_2\mathrm{E}(Y|X = c_2)\\ &= \mathrm{E}(Y).\end{aligned}$$

- Si $\mathrm{E}(Y|X = c_1) = \mathrm{E}(Y|X = c_2) = k$, alors leur moyenne vaudra k et donc $\mathrm{E}(Y) = k$.
- À l'inverse, $\mathrm{E}(Y) = k \not\Rightarrow \mathrm{E}(Y|X) = k$

Rappel : loi des espérances itérées (2/2)

On a de manière générale $\mathrm{E}(Y|X) = \mathrm{E}(Y)$ seulement si Y est sans corrélation avec les X. C'est le cas lorsque les variables sont indépendantes, et plus généralement sans corrélation (cf exemple du chapitre 1 où $\mathrm{corr}(X, Y) = 0$ mais où X et Y ne sont pas indépendantes) !

Principe général des variables instrumentales

$$Y = X\beta + \epsilon, \mathrm{E}(X'\epsilon) \neq 0.$$

Les erreurs sont corrélées avec les explicatives et les estimateurs habituels sont biaisés. L'idée est de trouver des variables Z, que l'on va appeler "instruments" telles que :

- $\mathrm{E}(Z'\epsilon) = 0$, les instruments sont sans corrélation avec le terme d'erreur. On dira alors qu'ils sont **valides**,
- $\mathrm{E}(Z'X) \neq 0$, les instruments sont corrélés avec les variables endogènes. S'ils ne le sont pas, on parlera d'instruments **faibles**.

Chacune de ces deux propriétés peut être testée (test de Sargan, de Stock et Yogo...). Dans la pratique, trouver des variables Z satisfaisant ces deux propriétés est difficile et demande souvent de tester des centaines d'ensembles de variables candidates...

Mise en oeuvre (1/2)

Des dizaines de possibilités existent : doubles moindres carrés (2SLS), Hausman et Taylor (1981), Arellano et Bond (1991), Blundell et Bond (1998)...

Les 2SLS restent la méthode la plus intuitive, constituée de deux étapes :

- dans une première étape, on régresse les variables explicatives endogènes sur les instruments et les variables explicatives exogènes. On calcule les valeurs prédites des endogènes,
- dans une seconde étape, on régresse Y sur les valeurs prédites des endogènes et les variables exogènes.

Intuitivement, on remplace les variables problématiques par des valeurs prédites avec les instruments. Si les instruments sont valides, les prévisions seront sans corrélation avec ϵ. Si les instruments sont faibles, on peut montrer que les résultats de deuxième étape seront peu précis et peu robustes (vraisemblance "plate" pour le modèle de la deuxième étape).

Mise en oeuvre (2/2)

Les subtilités sont nombreuses (correction de la matrice de variance lors de la deuxième étape, gestion des variables incluses/excluses, plusieurs tests possibles d'instruments faibles...). Dans le cas des variables instrumentales, le mieux est de ne jamais se lancer dans la programmation manuelle de ce type d'estimateurs et de toujours utiliser des **routines préprogrammées** et déjà testées. A ma connaissance, Stata est le seul logiciel à proposer des fonctions pour un grand nombre de modèles, qui fournissent de surcroit tous les critères et tests appropriés (routines `ivreg2` et `xtivreg2`, à installer manuellement).