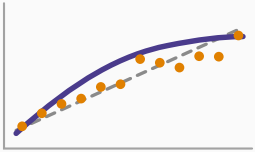
Économétrie avancée

Chapitre 9 : Problèmes de spécification et de données



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Plan du chapitre

Avant de commencer

Pourquoi ce chapitre ?

L'hétéroscédasticité (chapitre 8) est un problème d'**inférence** : les coefficients restent bien estimés. Ce chapitre traite un problème plus grave — la corrélation entre l'erreur u et un régresseur, qui biaise et rend incohérente l'estimation elle-même. Mauvaise forme fonctionnelle, variable omise, erreur de mesure, échantillon non aléatoire, observations aberrantes : autant de sources possibles, chacune avec son diagnostic et, parfois, son remède *via* les MCO.

Mauvaise spécification de la forme fonctionnelle

Diagnostiquer une forme fonctionnelle incorrecte

Un problème mineur en apparence, sérieux en pratique

Omettre $exper^2$ dans une équation de salaire, ou utiliser $wage$ plutôt que $\log(wage)$ quand ce dernier est la vraie forme, biaise généralement **tous** les coefficients — pas seulement celui de la variable mal spécifiée. Contrairement au cas d'une variable omise inobservable (section suivante), ici les données nécessaires sont disponibles : le problème est réparable en enrichissant la spécification.

Test RESET (Ramsey)

Le test **RESET** ajoute des puissances des valeurs ajustées $\hat{y}$ à l'équation initiale :

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + \text{erreur},$$

et teste $H_0 : \delta_1 = \delta_2 = 0$ par un F (approximativement $F_{2,\, n-k-3}$). Un rejet indique une non-linéarité négligée dans la relation entre y et les régresseurs — sans dire laquelle.

Prix immobiliers : niveau contre log-log

Sur les mêmes données, le modèle en niveau donne un RESET $= 4.67$ ($p = .012$, rejeté), tandis que le modèle log-log donne RESET $= 2.56$ ($p = .084$, non rejeté à 5 %). Le RESET oriente donc vers la spécification log-log — sans jamais garantir qu'elle est la forme "vraie".

Comparer des modèles non emboîtés

Test de Davidson-MacKinnon

Pour arbitrer entre $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ et sa variante log-log (non emboîtées, chapitre 6), on estime le modèle rival, on récupère ses valeurs ajustées $\tilde{y}$, et on teste la significativité de $\tilde{y}$ ajoutée au premier modèle : un t significatif rejette ce premier modèle. On peut appliquer le test symétriquement aux deux modèles — les deux peuvent être rejetés, ou aucun.

Variables proxy pour un facteur inobservé

Quand les données n'existent tout simplement pas

Contrairement à la mauvaise forme fonctionnelle, certains déterminants de y — l'aptitude innée dans une équation de salaire, la qualité de gestion dans une équation de profit — ne sont **jamais observés**. Omettre x_3^* corrélé à x_1 biaise $\hat{\beta}_1$ (chapitre 3); une **variable proxy**, corrélée à x_3^* sans lui être identique, peut atténuer ce biais.

Condition de validité d'une proxy

Avec $x_3^* = \delta_0 + \delta_3 x_3 + v_3$ reliant la variable inobservée x_3^* à sa proxy x_3, la solution "plug-in" (régresser y sur x_1, x_2, x_3) donne des estimateurs convergents de β_1, β_2 si $E(x_3^* \mid x_1, x_2, x_3) = E(x_3^* \mid x_3)$: une fois x_3 contrôlée, l'espérance de x_3^* ne dépend plus de x_1 ni x_2.

Le problème de la variable omise inobservable (suite)

Le QI comme proxy de l'aptitude

Dans une équation de $\log(wage)$, le rendement estimé de l'éducation tombe de 6,5 % à 5,4 % lorsqu'on ajoute le QI comme proxy de l'aptitude — cohérent avec un biais positif de variable omise (aptitude corrélée positivement à l'éducation). Le coefficient sur le QI lui-même ($+0{,}36\%$ par point) devient la quantité d'intérêt : une hausse d'un écart-type de QI (15 points) équivaut approximativement à une année d'éducation supplémentaire.

Variable dépendante retardée comme proxy

Contrôler l'historique inobservable

Inclure y_{t-1} (le taux de criminalité passé d'une ville, par exemple) dans une régression en coupe capture des facteurs historiques persistants et inobservés qui influencent à la fois y courant et une variable de politique publique (dépenses de sécurité). Comparer deux villes au même y_{t-1} et même taux de chômage isole mieux l'effet causal des dépenses — sans être une solution parfaite.

Application : criminalité et dépenses de sécurité

Sans le taux de criminalité passé, l'effet estimé des dépenses de sécurité sur la criminalité est faussement positif (villes à forte criminalité dépensent plus). En ajoutant $\log(crmrte_{t-5})$, l'élasticité des dépenses devient négative ($-0{,}14$) comme attendu — une correction de biais du même esprit que le contrôle par variables de panel (chapitres à venir).

Erreurs de mesure

Erreur de mesure sur la variable dépendante

Erreur sur y : généralement sans conséquence

Si $y = y^* + e_0$ où y^* est la variable d'intérêt réelle et e_0 l'erreur de mesure, et si e_0 est indépendante des régresseurs, les MCO sur y restent non biaisés et convergents — seule la variance de l'erreur augmente ($\sigma_u^2 + \sigma_{e_0}^2$), donc les erreurs-types sont plus grandes. Le biais n'apparaît que si e_0 est corrélée à un régresseur (ex : sous-déclaration systématique corrélée à une politique évaluée).

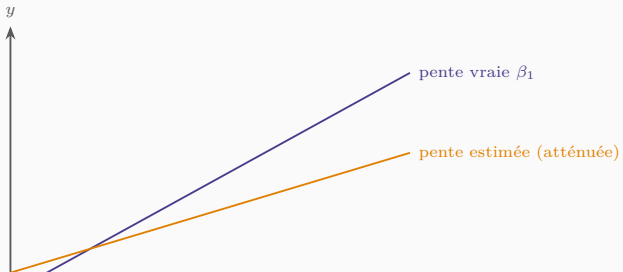
Erreur de mesure sur une variable explicative

Le modèle classique d'erreur de mesure (CEV)

Si $x_1 = x_1^* + e_1$ avec $\mathrm{Cov}(x_1^*, e_1) = 0$ (hypothèse CEV), alors x_1 et e_1 sont nécessairement corrélées, et régresser y sur x_1 produit un estimateur **biaisé vers zéro** :

$$\mathrm{plim}(\hat{\beta}_1) = \beta_1 \cdot \frac{\mathrm{Var}(x_1^*)}{\mathrm{Var}(x_1^*) + \mathrm{Var}(e_1)}.$$

C'est le **biais d'atténuation** : plus la variance de l'erreur de mesure est grande relative à celle du signal, plus $\hat{\beta}_1$ sous-estime $|\beta_1|$.



Données manquantes, échantillons non aléatoires, valeurs aberrantes

Aléatoire ou pas : toute la différence

Si l'absence de données est **indépendante** des variables du modèle, l'échantillon effectif reste aléatoire : on perd en précision, pas en validité. Si l'absence de donnée dépend elle-même de y ou de facteurs corrélés à u, l'échantillon devient non représentatif — c'est la **sélection non aléatoire**, source potentielle de biais.

Sélection exogène vs endogène

Une sélection fondée sur les **régresseurs** (par exemple, un échantillon restreint aux personnes de plus de 35 ans) laisse les MCO non biaisés — c'est une **sélection exogène**. Une sélection fondée sur la **variable dépendante** (ne garder que les ménages dont le patrimoine est inférieur à un seuil) biaise systématiquement l'estimation — c'est une **sélection endogène**, qui exige des méthodes dédiées (chapitres ultérieurs sur les données de panel et les modèles à variable latente).

Observations aberrantes

Pourquoi les MCO sont vulnérables

Les MCO minimisent une somme de carrés : un résidu deux fois plus grand pèse quatre fois plus dans l'ajustement. Une observation extrême — erreur de saisie, ou unité authentiquement atypique de la population — peut donc dominer l'estimation, surtout en petit échantillon.

Intensité de R&D et taille de firme

Sur 32 firmes chimiques, le coefficient de $sales$ dans une régression de $rdintens$ n'est pas significatif ; en retirant la seule firme dont les ventes dépassent 40 milliards de dollars (plus du double de toute autre), le coefficient triple et devient significatif ($t > 2$). Utiliser une forme log-log ($\log(rd)$ sur $\log(sales)$) réduit fortement la sensibilité à cette observation — un rappel du lien entre logarithmes et robustesse déjà noté au chapitre 6.

Observations aberrantes (suite)

Ne pas se fier au résidu seul pour repérer un outlier

Le résidu le plus grand en valeur absolue n'est pas nécessairement l'observation la plus influente : une observation à valeurs extrêmes sur les régresseurs peut avoir un petit résidu (elle "tire" la droite vers elle) tout en modifiant fortement les coefficients estimés. Toujours comparer les estimations avec et sans les observations suspectes plutôt que de se fier au seul diagnostic des résidus.

Une alternative robuste : les moindres écarts absolus (LAD)

L'estimateur **LAD** minimise $\sum_i |y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik}|$ plutôt que la somme des carrés — donnant moins de poids aux résidus extrêmes. Il est convergent vers l'espérance conditionnelle sous symétrie de l'erreur, mais moins efficace que les MCO si l'erreur est réellement normale, et nécessite une résolution numérique itérative.

Synthèse

Ce qu'il faut retenir

L'essentiel du chapitre 9

- La mauvaise forme fonctionnelle biaise les MCO mais se corrige avec les données disponibles ; le RESET la détecte sans l'identifier précisément.
- Une variable proxy corrige (partiellement) l'omission d'un facteur inobservé, sous la condition que son espérance conditionnelle isole correctement le facteur ciblé.
- L'erreur de mesure sur y est généralement inoffensive ; l'erreur de mesure classique (CEV) sur un régresseur biaise systématiquement vers zéro (atténuation).
- La sélection d'échantillon fondée sur les régresseurs est sans danger ; fondée sur y, elle biaise l'estimation. Les observations aberrantes doivent être diagnostiquées explicitement (comparaison avec/sans), pas seulement via les résidus.

Vers le chapitre 10

Jusqu'ici, chaque observation était tirée indépendamment des autres. Le chapitre suivant introduit les **séries temporelles** — des données où l'ordre chronologique compte, où les observations successives sont corrélées entre elles, et où de nouvelles hypothèses doivent remplacer l'échantillonnage aléatoire.