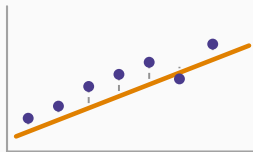


Introduction à l'économétrie

Chapitre 17 : La variable omise



Jaouad Madkour

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

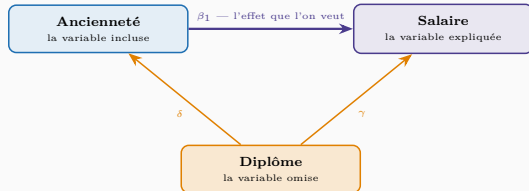
Le seul problème qui biaise

Prévoir le sens du biais

Les remèdes

Ce qu'il faut retenir

Le seul problème qui biaise



La formule du biais

$$\underbrace{\hat{\beta}_1^{\text{simple}}}_{277,17} = \underbrace{\beta_1}_{270,27} + \underbrace{\gamma \delta}_{\text{le biais}}$$

Le biais est le produit de deux effets : celui de la variable omise sur y , et sa liaison avec la variable incluse.

La vérification, sur nos données

$$1\,099,57 \times (-0,002907) = -3,20$$

$$3\,336,99 \times (+0,003025) = +10,09$$

$$\text{total} = +6,898$$

et $277,1674 - 270,2694 = 6,898$. La formule ne prédit pas approximativement : elle prédit *exactement*.

Deux conditions, et il faut les deux pour qu'il y ait biais.

La variable omise doit **influencer** y ($\gamma \neq 0$) et **être liée** à la variable incluse ($\delta \neq 0$). Si l'une des deux manque, l'omission est sans conséquence sur le coefficient.

Ici le biais est minuscule — sept dirhams sur deux cent soixante-dix — parce que δ est presque nul : dans cette entreprise, ancienneté et diplôme sont pratiquement indépendants. **Le danger vient des cas où ils ne le sont pas** : régresser le salaire sur la seule expérience en oubliant le diplôme, dans une population où les deux vont de pair, donnerait un coefficient massivement surestimé.

Et rien, dans la sortie du logiciel, ne l'aurait signalé.

Le biais de variable omise

Si le vrai modèle est $y = \beta_0 + \beta_1 x + \gamma z + \varepsilon$ et qu'on estime y sur x seul :

$$E(\hat{\beta}_1^{\text{simple}}) = \beta_1 + \gamma \delta$$

où δ est la pente de la régression de z sur x .

Elle se lit en une phrase

Le biais est le produit de **l'effet de la variable omise sur y** et de **sa liaison avec la variable incluse**.

Il faut les deux. Si la variable omise n'influence pas y , ou si elle est indépendante de x , l'omission est sans conséquence sur le coefficient.

Sur nos salaires

Coefficient de l'ancienneté, sans le diplôme : 277,1674. Avec le diplôme : 270,2694.

$$1\,099,5705 \times (-0,002907) = -3,197$$

$$3\,336,9916 \times (+0,003025) = +10,095$$

$$\text{total} = +6,898$$

Et l'écart observé vaut exactement 6,898

$$277,1674 - 270,2694 = 6,898.$$

La formule ne prédit pas approximativement : elle prédit exactement. Le biais de variable omise n'est pas une notion floue — c'est une identité algébrique, calculable dès qu'on dispose des deux modèles.

Parce que la liaison est presque nulle

Dans cette entreprise, ancienneté et diplôme sont pratiquement indépendants : δ vaut trois millièmes.

Sept dirhams de biais sur deux cent soixante-dix : négligeable. **Mais c'est une propriété de ces données-là, pas une règle générale.**

Le cas dangereux

Régresser le salaire sur la seule **expérience** dans une population où les plus expérimentés sont aussi les plus diplômés donnerait un coefficient massivement surestimé.

L'expérience capterait alors l'effet du diplôme. Et **rien, dans la sortie du logiciel, ne l'aurait signalé** : le R^2 serait honorable, le t écrasant, les résidus impeccables.

Prévoir le sens du biais

Le tableau des signes

γ	δ	Biais
+	+	surestimation
+	-	sous-estimation
-	+	sous-estimation
-	-	surestimation

Un raisonnement d'économiste, sans données

« J'ai omis les capacités individuelles. Elles augmentent le salaire ($\gamma > 0$) et vont de pair avec le diplôme ($\delta > 0$). Donc mon coefficient du diplôme est **surestimé**. »

On ne connaît pas l'ampleur du biais — mais on en connaît le **sens**, et cela suffit souvent à qualifier le résultat : « le rendement du diplôme est **au plus** de tant ».

Ce qui distingue ce problème de tous les autres

Trois différences

1. **Aucun test ne le détecte** — la variable manquante n'est pas dans les données.
2. **Aucune formule ne le répare** — ce n'est pas une correction d'écart-type.
3. **Il se raisonne, il ne se calcule pas** — sauf à disposer de la variable, auquel cas il suffit de l'ajouter.

La conséquence sur la façon de travailler

Un travail économétrique sérieux consacre plus de temps à la **spécification** qu'à l'estimation.

Choisir les variables, justifier leur présence, énumérer celles qui manquent et en discuter le sens : c'est cela, le métier. **L'estimation, elle, prend une ligne de code.**

Les remèdes

Par ordre de préférence

1. **Ajouter la variable**, si elle existe dans les données.
2. Trouver un **substitut** observable — le diplôme des parents pour le milieu social.
3. Utiliser une structure de données qui l'élimine : **panel**, effets fixes.
4. **Instrumenter** : trouver une variable liée à x mais pas à ε .

Les deux dernières dépassent ce cours

Les données de panel permettent d'éliminer tout ce qui est constant pour un individu — les capacités, par exemple — sans jamais l'observer.

Les variables instrumentales sont la réponse générale, et le chapitre 20 en dira un mot. **Les deux font l'objet de l'économétrie de Master.**

Le dire

Un résultat honnête accompagné de ses limites vaut mieux qu'un résultat présenté comme causal alors qu'il ne l'est pas.

La formulation correcte existe : « le diplôme est **associé** à un salaire supérieur de tant, sans qu'on puisse exclure que les capacités individuelles y contribuent ».

C'est la différence entre deux métiers

Celui qui écrit « le diplôme **augmente** le salaire de 44 % » a dépassé ce que ses données autorisent.

Celui qui écrit « les diplômés gagnent 44 % de plus, à ancienneté et région comparables » dit exactement ce qu'il a mesuré. **La seconde phrase est plus faible et plus solide.**

Ce qu'il faut retenir

Six points

1. C'est la **seule** violation qui biaise les coefficients eux-mêmes.
2. Biais = $\gamma \delta$: effet de la variable omise sur y , fois sa liaison avec x .
3. **Il faut les deux** : sans liaison, l'omission est sans conséquence.
4. Vérification exacte sur nos données : +6,898 prédits, +6,898 observés.
5. On peut souvent en **prévoir le sens** par le raisonnement économique, sans données.
6. Aucun test ne le détecte, aucune formule ne le répare : **c'est un problème de spécification**.

La suite

Il reste deux diagnostics à passer en revue — la normalité des résidus et les observations qui pèsent trop lourd.

Le chapitre 18 les traite ensemble, car ils se lisent sur les mêmes graphiques.