

# Biais par variables omises

Vincent Arel-Bundock



# Biais par variables omises

Problème:

- Une relation observée entre deux variables peut être factice si elle est causée par une troisième variable ignorée par notre modèle statistique.
- On dit alors que notre estimé de l'effet causal souffre d'un "Biais par Variables Omises."





Est-ce que les pompiers détruisent?

# Biais par variables omises: Exemple

Corrélation positive:

1. Nombre de pompiers déployés sur le site d'un incendie
2. Dommages causés par cet incendie.

Lorsqu'il y a plus de pompiers, les flammes détruisent plus.

Est-ce que cela veut dire que les pompiers causent les dommages?

Évidemment, l'intervention des pompiers réduit les dommages causés par le feu.



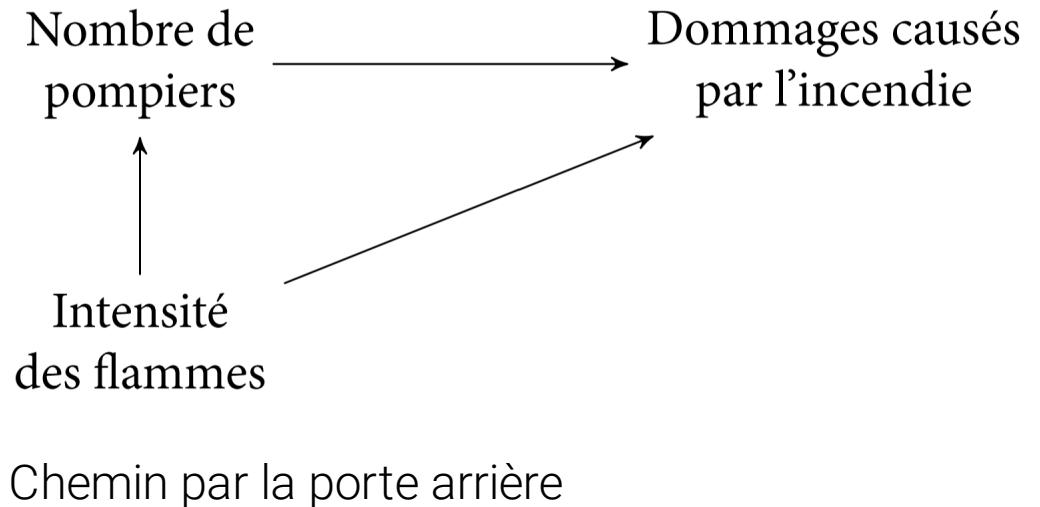
# Biais par variables omises

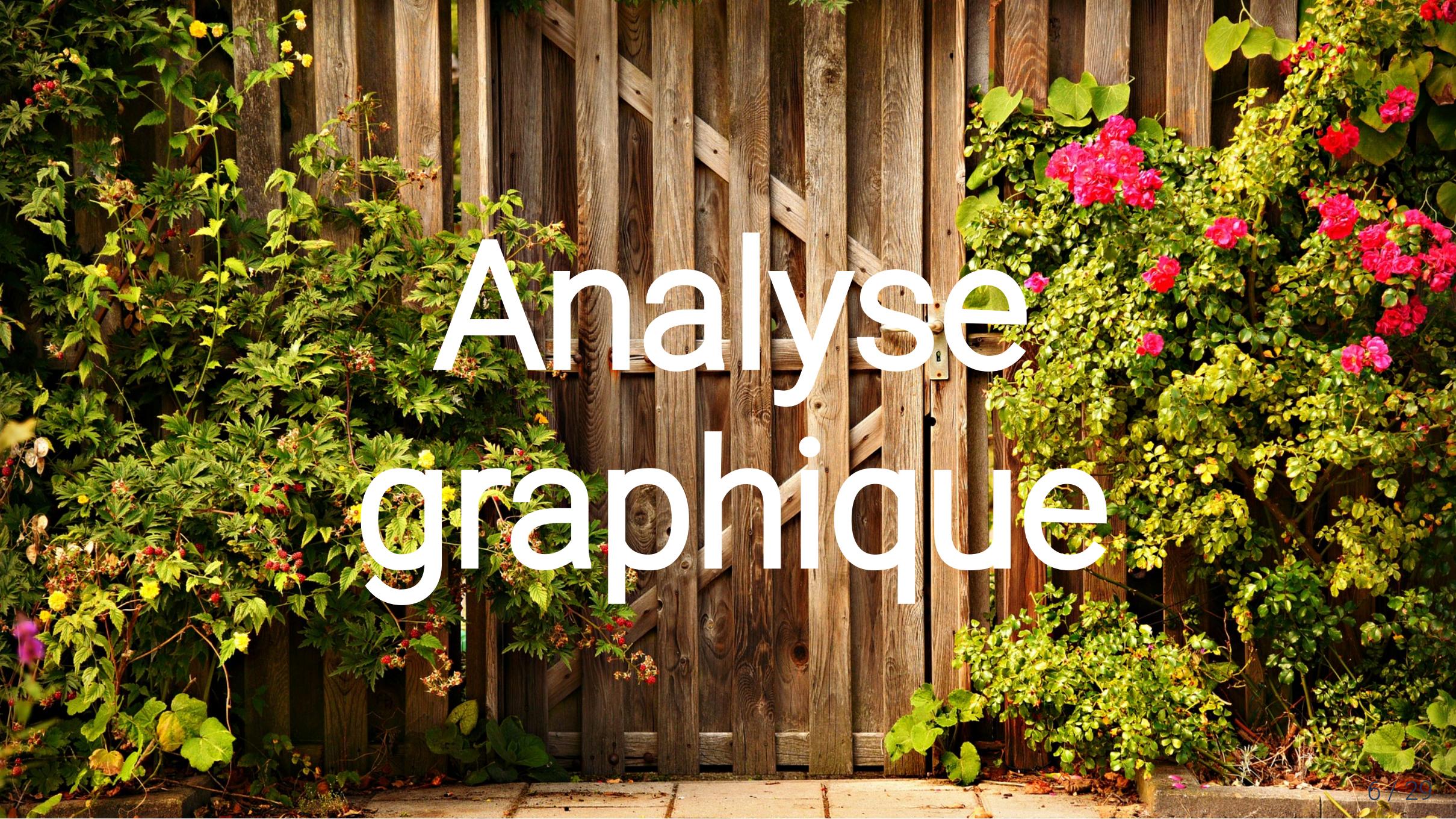
Variable omise:

- Intensité des flammes

Explication:

- Plus l'incendie est intense, plus il cause de dommages
- Plus l'incendie est intense, plus il faut de pompiers pour l'éteindre.



A photograph of a rustic wooden gate made of vertical planks. The gate is partially open, revealing a dark interior. It is surrounded by lush greenery, including ivy, roses, and yellow flowers, growing on both sides. The scene is bright and sunny, with shadows cast by the plants.

# Analyse graphique

# Biais par variables omises

Définition:

| Il y a un biais par variable omise si il existe un chemin ouvert par la porte arrière.

Condition #2 de l'identification causale dans la théorie causale structurelle.

# Exemple: Libres pour enfants

Plusieurs chercheurs ont observé une forte association entre le nombre de livres sur les étagères d'une maison, et le succès scolaire des enfants qui y habitent. Suggestion:

1. Augmenter le nombre de livres dans une maison améliore la performance des enfants qui y habitent;
2. Le gouvernement devrait mettre en place un programme massif d'achat de livres pour enfants.

Questions:

- Est-ce que distribuer des livres gratuits augmente la performance scolaire des enfants?
- Est-ce que ces conclusions sont justifiées?



Dolly Parton's Imagination Library

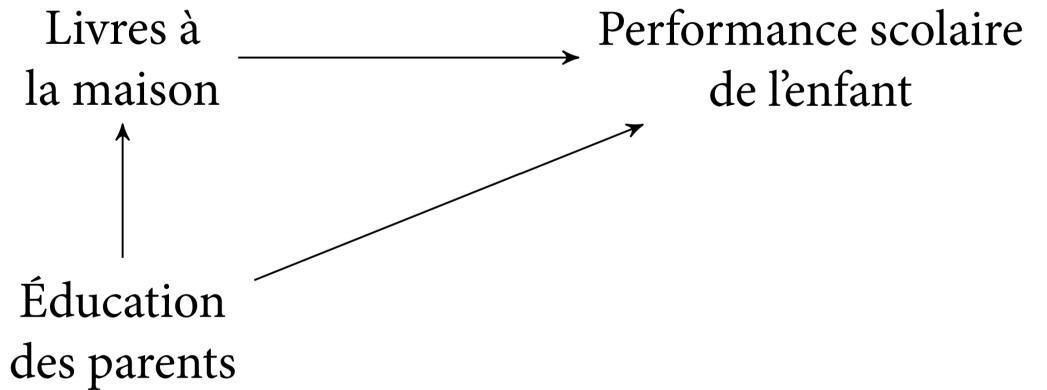
# Exemple: Livres pour enfants

Variable omise:

- Niveau d'éducation des parents.

Les parents plus éduqués achètent plus de livres, et ils transmettent des compétences scolaires à leurs enfants.

Chemin par la porte arrière



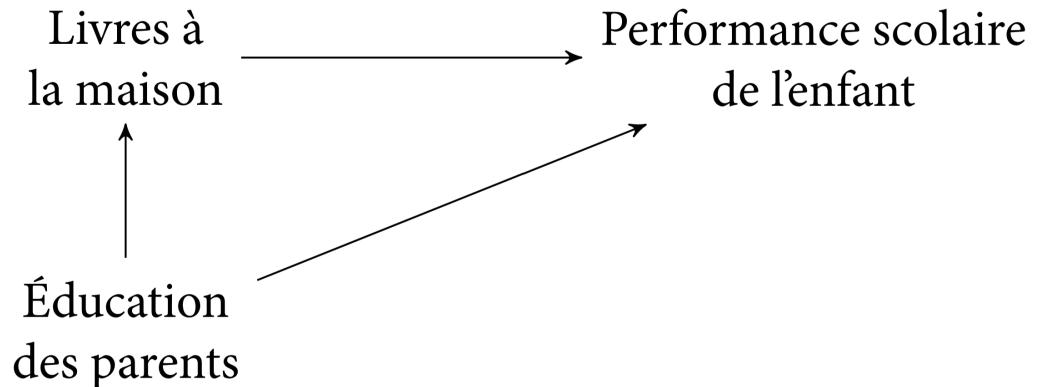
# Exemple: Livres pour enfants

Pour estimer l'effet du nombre de livres sur le succès scolaire, il faut bloquer le chemin par la porte arrière en contrôlant pour le niveau d'éducation des parents:

Livres ← **Éducation** → Performance

Étudier seulement l'association bivariée entre livres et performance, sans contrôler pour l'éducation des parents, produirait une conclusion erronée.

Si la vraie cause du succès scolaire est la transmission du savoir des parents à l'enfant, un programme gouvernemental d'achat de livres risque d'être inefficace.



# Exemple: Alcool et diabète

Deux études:

1. Holst et al. (2017) analysent un sondage mené auprès de 70000 résidents du Danemark, et estiment que la consommation de bière est associée à un risque réduit de diabète.
2. Griswold et al. (2018) concluent qu'il est plus sécuritaire de ne pas consommer d'alcool du tout.

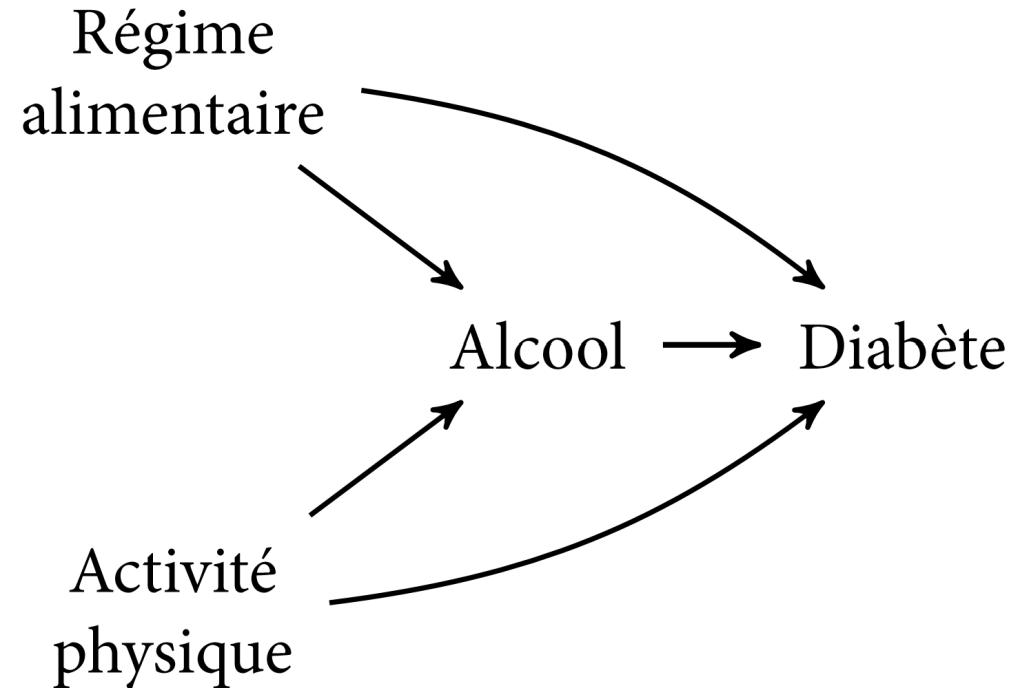
Causalité? Prudence!



# Exemple: Alcool et diabète

Les gens qui choisissent un régime alimentaire faible en calories pourraient éviter l'alcool, puisque les boissons alcoolisées ont souvent un haut contenu calorique.

Les sportifs pourraient s'abstenir de consommer de l'alcool pour éviter que leur performance soit affectée:



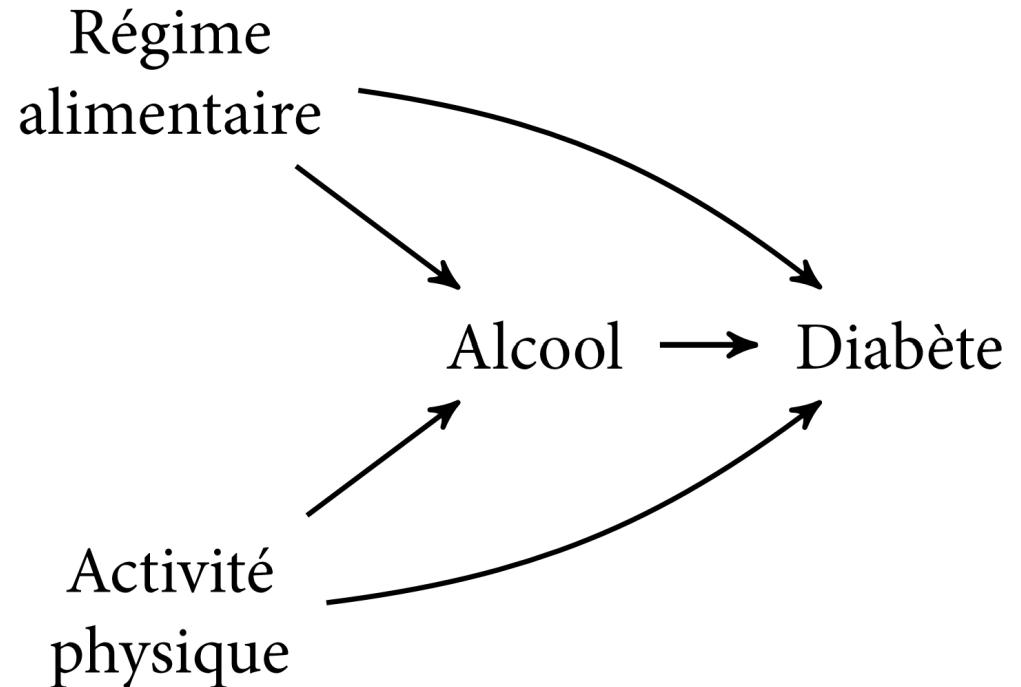
# Exemple: Alcool et diabète

Il y a deux sources de biais par variable omise causés par deux chemins par la porte arrière ouverts.

Pour obtenir un estimé non biaisé de l'effet de l'alcool sur la santé, il faut bloquer ces deux chemins:

$$\begin{aligned} \text{Alcool} &\leftarrow \boxed{\text{Régime alimentaire}} \rightarrow \text{Diabète} \\ \text{Alcool} &\leftarrow \boxed{\text{Activité physique}} \rightarrow \text{Diabète} \end{aligned}$$

Pour estimer l'effet causal de l'alcool sur la santé à l'aide de données d'observation (p.ex. un sondage), il faut identifier toutes les variables susceptibles d'introduire de tels biais par variable omise.



# Solution graphique

L'analyse graphique permet de déterminer si les estimés souffrent d'un biais par variable omise en vérifie que les deux conditions de l'identification causale soiesnt satisfaites :

1. Le modèle ne contrôle pas pour un descendant de la cause.
2. Tous les chemins par la porte arrière sont bloqués.

Dans une étude observationnelle, il est difficile de garantir que ces deux conditions sont remplies.

On peut au mieux identifier et contrôler pour les *principales* sources de biais par variable omise.

# Analyse algébrique

# Analyse algébrique

Analyse graphique:

- Conditions d'identification causale

Analyse algébrique:

- Conditions d'identification causale
- Anticiper la taille et la direction du biais

# Analyse algébrique

Objectif:

- Estimer l'effet causal de  $X$  sur  $Y$  à l'aide d'un modèle de régression "bivarié" ou "court".

$$Y = \alpha_0 + \alpha_1 X + \nu$$

L'estimé du coefficient  $\alpha_1$  est obtenu en appliquant cette formule:

$$\hat{\alpha}_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}$$

# Analyse algébrique

Objectif:

- Estimer l'effet causal de  $X$  sur  $Y$  à l'aide d'un modèle de régression "bivarié" ou "court".

$$Y = \alpha_0 + \alpha_1 X + \nu$$

L'estimé du coefficient  $\alpha_1$  est obtenu en appliquant cette formule:

$$\hat{\alpha}_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}$$

Variable omise: modèle court incomplet. Modèle "véridique" ou "complet":

$$Y = \beta_0 + \beta_1 X + \beta_2 A + \varepsilon$$

Questions:

1. Est-ce que  $\alpha_1 = \beta_1$
2. Est-ce que la variable  $A$  introduit un biais par variable omise dans l'estimé du coefficient  $\alpha_1$ ?

# Analyse algébrique

$$\begin{aligned}\hat{\alpha}_1 &= \frac{\text{Cov}(Y, X)}{\text{Var}(X)} \\ &= \frac{\text{Cov}(\beta_0 + \beta_1 X + \beta_2 A + \varepsilon, X)}{\text{Var}(X)} \\ &= \frac{\text{Cov}(\beta_0, X) + \text{Cov}(\beta_1 X, X) + \text{Cov}(\beta_2 A, X) + \text{Cov}(\varepsilon, X)}{\text{Var}(X)} \\ &= \frac{\text{Cov}(\beta_1 X, X) + \text{Cov}(\beta_2 A, X) + \text{Cov}(\varepsilon, X)}{\text{Var}(X)} \\ &= \frac{\beta_1 \text{Cov}(X, X) + \beta_2 \text{Cov}(A, X) + \text{Cov}(\varepsilon, X)}{\text{Var}(X)}\end{aligned}$$

# Analyse algébrique

$$\begin{aligned}\hat{\alpha}_1 &= \frac{\text{Cov}(Y, X)}{\text{Var}(X)} \\ &= \frac{\text{Cov}(\beta_0 + \beta_1 X + \beta_2 A + \varepsilon, X)}{\text{Var}(X)} \\ &= \frac{\text{Cov}(\beta_0, X) + \text{Cov}(\beta_1 X, X) + \text{Cov}(\beta_2 A, X) + \text{Cov}(\varepsilon, X)}{\text{Var}(X)} \\ &= \frac{\text{Cov}(\beta_1 X, X) + \text{Cov}(\beta_2 A, X) + \text{Cov}(\varepsilon, X)}{\text{Var}(X)} \\ &= \frac{\beta_1 \text{Cov}(X, X) + \beta_2 \text{Cov}(A, X) + \text{Cov}(\varepsilon, X)}{\text{Var}(X)}\end{aligned}$$

# Analyse algébrique

$$\hat{\alpha}_1 = \beta_1 + \beta_2 \cdot \frac{\text{Cov}(A, X)}{\text{Var}(X)} + \frac{\text{Cov}(\varepsilon, X)}{\text{Var}(X)}$$

Si le modèle multivarié est "véridique" ou "complet," cela signifie que  $X \perp \varepsilon$ . En moyenne, la covariance entre  $X$  et  $\varepsilon$  sera donc égale à zéro:

$$\hat{\alpha}_1 = \beta_1 + \beta_2 \cdot \frac{\text{Cov}(A, X)}{\text{Var}(X)}$$

Cette équation montre que les coefficients des modèles "court" et "long" ne sont pas égaux :

$$E[\hat{\alpha}_1] \neq E[\beta_1]$$

$\hat{\alpha}_1$  est biaisé.  $\hat{\alpha}_1$  n'est pas causal.

# Analyse algébrique

$$\hat{\alpha}_1 = \beta_1 + \beta_2 \cdot \frac{\text{Cov}(A, X)}{\text{Var}(X)} + \frac{\text{Cov}(\varepsilon, X)}{\text{Var}(X)}$$

Le biais de l'estimé  $\hat{\alpha}_1$  est égal à  $\beta_2 \cdot \frac{\text{Cov}(A, X)}{\text{Var}(X)}$ .

Fraction équivalente au coefficient de régression de ce modèle :

$$A = \pi_0 + \pi_1 X + \gamma, \quad \text{où } \pi_1 = \frac{\text{Cov}(A, X)}{\text{Var}(X)}$$

# Analyse algébrique

$$\hat{\alpha}_1 = \beta_1 + \beta_2 \cdot \frac{\text{Cov}(A, X)}{\text{Var}(X)} + \frac{\text{Cov}(\varepsilon, X)}{\text{Var}(X)}$$

Le biais de l'estimé  $\hat{\alpha}_1$  est égal à  $\beta_2 \cdot \frac{\text{Cov}(A, X)}{\text{Var}(X)}$ .

Fraction équivalente au coefficient de régression de ce modèle :

$$A = \pi_0 + \pi_1 X + \gamma, \quad \text{où } \pi_1 = \frac{\text{Cov}(A, X)}{\text{Var}(X)}$$

Les résultats obtenus jusqu'à maintenant peuvent être résumés ainsi :

$$\underbrace{\hat{\alpha}_1}_{\text{Estimé}} = \overbrace{\beta_1}^{\text{Vérité}} + \underbrace{\beta_2 \cdot \pi_1}_{\text{Biais}}$$

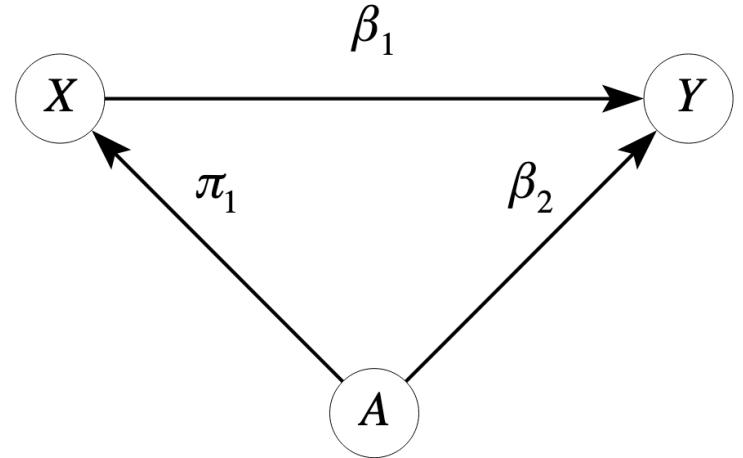
# Analyse algébrique

$$\underbrace{\hat{\alpha}_1}_{\text{Estimé}} = \overbrace{\beta_1}^{\text{Vérité}} + \underbrace{\beta_2 \cdot \pi_1}_{\text{Biais}}$$

Le biais par variable omise qui affecte le modèle "court" dépend de deux facteurs:

1.  $\beta_2$ : La relation entre la variable omise ( $A$ ) et la variable dépendante ( $Y$ ).
2.  $\pi_1$ : La relation entre la variable omise ( $A$ ) et la variable indépendante ( $X$ ).

Si l'un ou l'autre de ces coefficients est égal à zéro, nous pourrions estimer le modèle "court" et ignorer la variable  $A$  sans craindre que nos résultats soient biaisés.



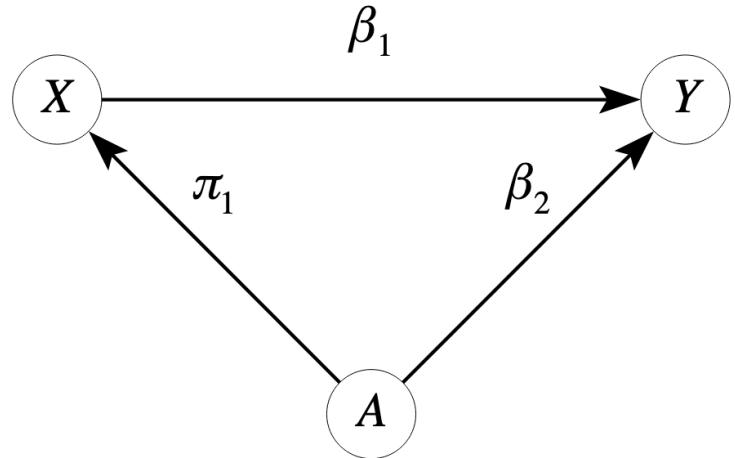
# Analyse algébrique

$$\underbrace{\hat{\alpha}_1}_{\text{Estimé}} = \overbrace{\beta_1}^{\text{Vérité}} + \underbrace{\beta_2 \cdot \pi_1}_{\text{Biais}}$$

Cette équation nous informe sur la direction et la force du biais par variable omise :

- La *direction* dépend du signe des deux relations.
- La *taille* dépend de la force des deux relations.

Plus les relations  $\beta_2$  et  $\pi_1$  sont fortes, plus le biais risque d'être important.



# Analyse algébrique

$$\underbrace{\hat{\alpha}_1}_{\text{Estimé}} = \overbrace{\beta_1}^{\text{Vérité}} + \underbrace{\beta_2 \cdot \pi_1}_{\text{Biais}}$$

$$Y = \alpha_0 + \alpha_1 X + \nu$$

$$Y = \beta_0 + \beta_1 X + \beta_2 A + \varepsilon$$

$$A = \pi_0 + \pi_1 X + \gamma, \quad \text{ où } \pi_1 = \frac{\text{Cov}(A, X)}{\text{Var}(X)}$$

		Relation entre $A$ et $Y$	
		+	-
Relation entre $A$ et $X$	+	Biais positif	Biais négatif
	-	Biais négatif	Biais positif

# Exemple: Livres et performance scolaire

Pour étudier l'effet causal des livres sur la performance scolaire des enfants, on pourrait estimer un modèle court, un modèle long, et un modèle auxiliaire :

$$\text{Performance} = \alpha_0 + \alpha_1 \text{Livres} + \nu$$

$$\text{Performance} = \beta_0 + \beta_1 \text{Livres} + \beta_2 \text{Éducation des parents} + \varepsilon$$

$$\text{Éducation des parents} = \pi_0 + \pi_1 \text{Livres} + \gamma$$

Le premier modèle risque d'offrir un estimé biaisé de l'effet des livres sur la performance scolaire ( $\alpha_1$ ).

Le coefficient de régression du modèle incomplet est égal à:

$$\alpha_1 = \beta_1 + \beta_2 \cdot \pi_1$$

# Exemple: Livres et performance scolaire

$$\text{Performance} = \alpha_0 + \alpha_1 \text{Livres} + \nu$$

$$\begin{aligned}\text{Performance} &= \beta_0 + \beta_1 \text{Livres} + \beta_2 \text{Éducation des parents} + \varepsilon \\ \text{Éducation des parents} &= \pi_0 + \pi_1 \text{Livres} + \gamma\end{aligned}$$

$$\alpha_1 = \beta_1 + \beta_2 \cdot \pi_1$$

Si l'éducation des parents est positivement associée à la performance scolaire des étudiants ( $\beta_2 > 0$ ), et si l'éducation des parents est positivement associée au nombre de livres disponibles ( $\pi_1 > 0$ ), alors le biais par variable omise est positif ( $\beta_2 \cdot \pi_1 > 0$ ).

Estimer un modèle naïf bivarié tend à *surestimer* l'effet positif des livres sur la performance des étudiants :

- Plus la relation entre l'éducation des parents et la performance scolaire est forte, plus  $\alpha_1$  risque d'être biaisé.
- Plus la relation entre l'éducation des parents et le nombre de livres à la maison est forte, plus  $\alpha_1$  risque d'être biaisé.

# Exemple: Alcool et santé

Un chercheur qui s'intéresse à l'effet de l'alcool sur la santé pourrait estimer trois modèles:

$$\text{Santé} = \alpha_0 + \alpha_1 \text{Alcool} + \nu$$

$$\text{Santé} = \beta_0 + \beta_1 \text{Alcool} + \beta_2 \text{Exercice} + \varepsilon$$

$$\text{Exercice} = \pi_0 + \pi_1 \text{Alcool} + \gamma$$

Le premier de ces trois modèles risque d'offrir un estimé biaisé des bienfaits de l'alcool pour la santé ( $\alpha_1$ ).

Si l'exercice physique est lié à un meilleur état de santé ( $\beta_2 > 0$ ) et si la consommation d'alcool est négativement associée à l'exercice physique ( $\pi_1 < 0$ ), alors le biais par variable omise est négatif:  $\beta_2 \cdot \pi_1 < 0$ .

Estimer un modèle naïf bivarié tend à sous-estimer les bienfaits de la consommation d'alcool pour la santé (ou à exagérer ses méfaits).

The background of the image is a vast, clear blue sky filled with large, white, puffy cumulus clouds. The clouds are scattered across the frame, with some appearing in the foreground and others in the middle ground. The lighting suggests a bright, sunny day.

Limite

# Limite de l'approche algébrique

$$\underbrace{\hat{\alpha}_1}_{\text{Estimé}} = \overbrace{\beta_1}^{\text{Vérité}} + \underbrace{\beta_2 \cdot \pi_1}_{\text{Biais}}$$

Cette équation mesure le biais qui survient dans un modèle avec seulement deux variables explicatives.

En pratique, il y a souvent beaucoup de variables explicatives, et nous pouvons rarement toutes les inclure dans nos modèles de régression.

Si plusieurs variables introduisent des biais de différentes tailles et de différents signes, il devient difficile d'anticiper la taille ou la direction du biais total.

L'équation est une règle approximative plutôt qu'une loi déterminante.

# Merci!

