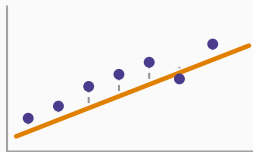


Introduction à l'économétrie

Chapitre 20 : Choisir un modèle, et le seuil de la causalité



Jaouad Madkour

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

Choisir

Le seuil de la causalité

La conclusion du cours

Ce qu'il faut retenir

Choisir

Deux erreurs opposées

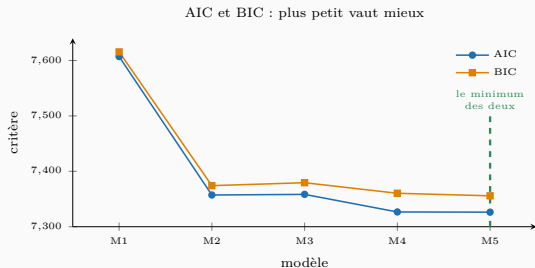
Trop peu de variables : risque de **variable omise**, donc de biais.

Trop de variables : perte de degrés de liberté, multicolinéarité, et un modèle qui décrit l'échantillon au lieu de la population.

Les deux erreurs ne se valent pas

Ajouter une variable inutile coûte de la **précision**. Omettre une variable pertinente coûte de la **justesse**.

En cas de doute, mieux vaut inclure. C'est le même arbitrage qu'au chapitre 14 : on ne troque pas un problème de variance contre un problème de biais.



Les cinq modèles

M1 ancienneté seule

M2 + diplôme

M3 + sexe

M4 + région

M5 = M4 *sans* le sexe

$$AIC = n \ln\left(\frac{SCR}{n}\right) + 2p$$

$$BIC = n \ln\left(\frac{SCR}{n}\right) + p \ln n$$

Les deux ajoutent une **pénalité** au manque d'ajustement. Le BIC pénalise plus lourdement : il choisit des modèles plus courts.

Ici les deux désignent M5, le modèle sans le sexe. La variable non significative est écartée par le critère, sans qu'on ait eu à trancher à la main.

La structure commune

$$\text{critère} = \underbrace{n \ln\left(\frac{\text{SCR}}{n}\right)}_{\text{mauvais ajustement}} + \underbrace{\text{pénalité} \times p}_{\text{complexité}}$$

AIC pénalise de 2 par paramètre; BIC de $\ln n$, soit 6,2 ici. **Le BIC est donc plus sévère et choisit des modèles plus courts.**

Sur nos données, les deux s'accordent

Le minimum est atteint par **M5** — le modèle avec ancienneté, diplôme et région, **sans le sexe**.

La variable non significative est écartée par le critère, sans qu'on ait eu à trancher à la main. Les tests, les critères et le R^2 ajusté disent tous la même chose.

Une variable théoriquement indispensable se garde

Si la théorie économique impose qu'une variable figure dans le modèle, elle y reste — même non significative, même si le critère préfère s'en passer.

Un modèle de demande sans le prix n'est pas un modèle de demande, quel que soit son AIC.

Et la sélection automatique est une mauvaise idée

Les procédures pas à pas — ajouter ou retirer selon les t — enchaînent des dizaines de tests et invalident tous les seuils, exactement comme la pêche aux résultats du chapitre 15 de l'inférence.

Le modèle final n'a alors plus de p -values interprétables, puisqu'il a été choisi *sur* les données qui servent à le tester.

Le seuil de la causalité

La question que tout le monde pose



La variable confondante
elle explique les deux, et fausse β

La causalité inverse
le budget publicitaire suit les ventes passées

Dans les deux cas, la variable explicative est corrélée au terme d'erreur : l'hypothèse H2 tombe, et le coefficient est biaisé.

C'est le mot d'**endogénéité**, et c'est le seuil où s'arrête ce cours. Les capacités non observées gonflent le rendement apparent du diplôme ; les ventes qui gonflent le budget publicitaire font surestimer l'effet de la publicité.

Aucun test ne le révèle, aucune correction d'écarts-types n'y remédie. La sortie du logiciel est parfaitement propre — et le coefficient reste faux.

Les réponses existent : variables instrumentales, doubles différences, discontinuité de régression, expériences contrôlées. Elles font l'objet de l'économétrie de niveau Master, et c'est le sujet qui a valu plusieurs prix Nobel depuis 2000. **Retenez seulement, de ce cours, que « significatif » ne veut jamais dire « causal ».**

Ce qu'un coefficient établit

Trois formulations, une seule correcte

Formulation	Statut
« Les diplômés gagnent 44 % de plus, à ancienneté comparable »	correcte
« Le diplôme est associé à 44 % de salaire en plus »	correcte
« Le diplôme augmente le salaire de 44 % »	dépasse les données

La troisième suppose ce qu'on n'a pas montré

Elle suppose que si l'on donnait un Bac+5 à un salarié pris au hasard, son salaire monterait de 44 %.

Or les diplômés diffèrent des autres par bien d'autres choses — capacités, milieu, réseau, ambition — dont aucune n'est dans le fichier. **Une partie de l'écart leur revient.**

Endogénéité

La variable explicative est corrélée au terme d'erreur : $\text{cov}(x, \varepsilon) \neq 0$.

L'hypothèse H2 tombe, et le coefficient est **biaisé**.

Ses trois sources

1. **Variable omise** — chapitre 17.
2. **Causalité inverse** — le budget publicitaire suit les ventes autant qu'il les précède.
3. **Erreur de mesure** sur la variable explicative.

Et ne pouvait pas résoudre

Aucun test ne détecte l'endogénéité. Aucune correction d'écarts-types n'y remédie. La sortie du logiciel reste parfaitement propre, et le coefficient reste faux.

C'est la limite de tout ce qui précède, et il vaut mieux la connaître que la découvrir dans un rapport publié.

Les réponses existent, et elles sont récentes

Variables instrumentales, doubles différences, discontinuité de régression, expériences contrôlées : ce sont les méthodes de l'**inférence causale**.

Elles font l'objet de l'économétrie de Master, et plusieurs prix Nobel d'économie leur ont été décernés depuis 2000. Toutes reposent sur une idée simple : **trouver, dans les données, quelque chose qui ressemble à une expérience**.

La conclusion du cours

Trois phrases, et elles suffisent

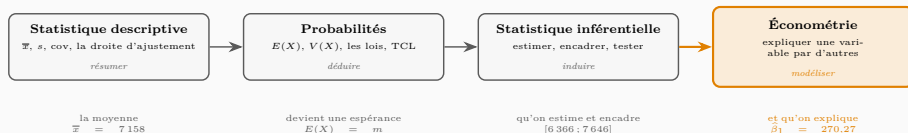
1. « Voici ce que j'ai mesuré, et avec quelle précision. »
2. « Voici les hypothèses sous lesquelles cette mesure est valide. »
3. « Voici ce que je ne peux pas conclure. »

La troisième est celle qui distingue le professionnel

Tout le monde sait produire une régression. Peu savent en énoncer les limites — et ce sont les seuls dont le travail résiste à l'examen.

« Significatif » ne veut jamais dire « causal ». Si vous ne deviez retenir qu'une phrase de ce cours, ce serait celle-là.

Ce que quatre cours ont construit



Un seul jeu de données a traversé les quatre cours : les mêmes cinq cents salariés.

La statistique descriptive en a tiré une moyenne, un écart-type et une droite d'ajustement. Les probabilités ont transformé la fréquence en probabilité et la moyenne en espérance. L'inférence a fait de ces cinq cents individus une population dont on tire des échantillons, pour estimer, encadrer et tester. L'économétrie, enfin, a posé qu'il existe une *relation vraie* entre le salaire, l'ancienneté, le diplôme et la région — et l'a mesurée.

Aucun de ces quatre cours n'a introduit d'outil que les précédents n'annonçaient pas. La covariance de la descriptive est devenue la pente ; le théorème central limite des probabilités a rendu les tests possibles ; l'erreur-type de l'inférence est devenue celle d'un coefficient. C'est une seule discipline, enseignée en quatre fois.

Ce qu'il faut retenir

Six points

1. Trop peu de variables biaise; trop de variables imprécise. **En cas de doute, inclure.**
2. **AIC et BIC** ajoutent une pénalité de complexité; le BIC est plus sévère.
3. Sur nos données, les deux choisissent **M5** — sans le sexe.
4. Une variable **théoriquement indispensable** se garde, quel que soit le critère.
5. La **sélection automatique** invalide les p -values du modèle final.
6. **Endogénéité** : variable omise, causalité inverse, erreur de mesure. Aucun test, aucune correction.

La suite

Tout le cours a été conduit à la main, pour que chaque nombre soit vérifiable.

Les deux derniers chapitres le refont sur machine : une étude de cas complète, puis la mise en œuvre sous R et Python.