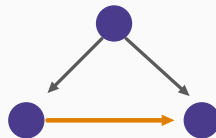


Inférence causale et machine learning

Chapitre 1 : Corrélation et causalité : ce que la régression ne dit pas



Prérequis

La question de ce cours

Le piège élémentaire

Pourquoi contrôler ne suffit pas

Le plan du cours

Ce qu'il faut retenir

Prérequis

Les cours qui précèdent

Économétrie avancée et *Apprentissage statistique*.

Les notions mobilisées

De l'économétrie : les variables instrumentales, les données de panel, le biais de variable omise.

De l'apprentissage : la validation croisée, la régularisation, et l'idée qu'un modèle se juge hors échantillon.

Ce que ce cours ajoute

L'économétrie sait estimer une corrélation conditionnelle ; l'apprentissage sait prévoir. **Ni l'un ni l'autre ne répond à la question de l'économiste** : que se passerait-il si l'on changeait la politique ?

Ce cours ferme la branche en revenant à cette question, avec les outils des deux précédents.

La question de ce cours

Ce que les deux cours précédents ont laissé de côté

Le point d'arrivée d'Apprentissage statistique

Ce cours-là s'achevait sur un avertissement répété : le lasso sélectionne pour prévoir, l'importance des variables n'est pas un effet, une valeur de Shapley explique le **modèle** et non le **monde**.

Il concluait qu'aucune méthode d'interprétation ne répond à « quel serait l'effet si l'on intervenait ? ».

Ce cours prend cette question en charge.

Deux questions, irréductibles l'une à l'autre

Question	Nature	Outil
Quelle sera la valeur de Y ?	prévision	apprentissage statistique
Que se passerait-il si l'on changeait T ?	causalité	ce cours

Application en gestion

Un modèle qui prévoit le défaut à partir du nombre de relances est excellent. Il ne dit pas s'il faut cesser de relancer.

Le décideur ne choisit pas ce qu'il observe : il choisit ce qu'il **fait**. Toute évaluation de politique — un salaire minimum, une formation, un dispositif fiscal — est une question causale, et aucune performance prédictive n'y répond.

Le piège élémentaire

Un exemple qui suffit à convaincre

Un programme de formation, évalué naïvement

Mille personnes formées, mille non formées. On observe le taux d'emploi un an après.

	Formés	Non formés
Taux d'emploi	$400/1\,000 = 40\%$	$600/1\,000 = 60\%$

La formation semble nuire, et lourdement : vingt points d'écart.

Le même tableau, par niveau de qualification

Groupe	Formés	Non formés
Peu qualifiés	$240/800 = \mathbf{30\%}$	$40/200 = 20\%$
Très qualifiés	$160/200 = \mathbf{80\%}$	$560/800 = 70\%$
Ensemble	$\mathbf{40\%}$	$\mathbf{60\%}$

Pourquoi contrôler ne suffit pas

Ajouter des variables ne règle pas la question

1. On ne peut contrôler que ce qu'on **observe** — la motivation, l'aptitude, l'information privée échappent.
2. Certaines variables **ne doivent pas** être contrôlées : le chapitre 3 montrera qu'en ajouter peut créer un biais.
3. Rien dans les données ne dit **lesquelles** contrôler : c'est une décision qui vient d'ailleurs.

Le réflexe à désapprendre

« J'ai mis tous les contrôles disponibles » n'est pas un argument d'identification. C'est souvent l'inverse : une régression saturée de variables mal choisies produit des coefficients dont personne ne sait ce qu'ils mesurent.

La qualité d'une estimation causale ne se lit pas dans le R^2 , ni dans le nombre de contrôles, ni dans la significativité. Elle se lit dans l'argument qui justifie la comparaison.

Intuition

Il faut un langage pour écrire ce qu'on **voudrait** observer avant d'écrire ce qu'on estime — et des conditions explicites sous lesquelles ce qu'on estime correspond à ce qu'on veut.

C'est le cadre des résultats potentiels, au chapitre 2.

Le plan du cours

Trois mouvements

Ce que traitent les douze chapitres

Chapitres	Objet
1–4	Le cadre : résultats potentiels, graphes causaux, expérience aléatoire
5–9	Les stratégies d'identification quand on ne peut pas randomiser
10–12	Ce que le machine learning apporte, et la décision

Le fil rouge

Le cours suit une même question : **quel est l'effet d'un programme de formation sur l'emploi et le salaire ?**

Chaque méthode y sera appliquée, avec ses hypothèses et ses limites. Comparer les réponses obtenues est plus instructif que d'accumuler des cas différents.

Ce qu'il faut retenir

Cinq points

1. **Prévoir** et **estimer un effet** sont deux questions irréductibles.
2. Le **paradoxe de Simpson** : le traitement gagne dans chaque groupe et perd au total.
3. Le problème est d'**identification**, non d'échantillon : plus de données ne corrige rien.
4. **Contrôler ne suffit pas** — on ne contrôle que l'observable, et certains contrôles nuisent.
5. La crédibilité vient de l'**argument**, pas du R^2 ni du nombre de variables.

La suite

Le chapitre 2 introduit la notation qui rend tout cela précis, et qui permet de dire exactement ce qui manque.