Résumé structurel du cours d'Économétrie 1

(Lucas Girard) – Cette version : 15 janvier 2024

Version française (version anglaise sur Pamplemousse)

Préambule *Ces quelques rappels ne sont pas exhaustifs, ils ne reprennent pas tous les éléments au programme de l'examen ; ils cherchent plutôt à insister sur les principales articulations du cours, sa structure, afin de faciliter votre compréhension globale du cours, pour vous aider pour l'examen et peut-être même davantage et surtout pour votre compréhension et mémorisation à plus long terme des notions et réflexes importants vus dans ce cours.*

Notation Les notations utilisées reprennent celles du cours (voir au besoin les rappels au début des quiz). Puisqu'on suppose toujours les échantillons indépendants et identiquement distribués (i.i.d), on rappelle qu'on omet l'indice i pour désigner une instance générique ayant la même distribution que les données. Par exemple, $(Y, D) \sim \mathrm{P}_{(Y,D)}$ désigne une instance générique du couple résultat $\times$ traitement ayant la même distribution, notée $\mathrm{P}_{(Y,D)}$, que celle d'une observation quelconque (Y_i, D_i), $i \in \{1, \dots, n\}$, avec n la taille de l'échantillon.
Il est primordial de toujours bien avoir en tête *la nature* et *la dimension* des objets considérés.
Par défaut, les vecteurs (aléatoires ou non) sont ici des matrices colonnes et $\cdot'$ désigne la transposée.
Pour deux ensembles E et F, F^E désigne l'ensemble des applications de E dans F.
On utilise cette notation pour bien distinguer les variables ou vecteurs aléatoires / stochastiques et les paramètres non stochastiques. Par exemple, l'effet causal individuel $\Delta \in \mathbb{R}^\Omega$ est une variable aléatoire réelle, c'est-à-dire une application mesurable de Ω dans $\mathbb{R}$, où $(\Omega, \mathcal{A}, \mathbb{P})$ est l'espace probabilisé sous-jacent depuis lequel sont définies toutes les variables aléatoires considérées ; par contre $\delta := \mathbb{E}[\Delta] \in \mathbb{R}$, l'effet causal moyen, est un nombre réel non stochastique (le cadre est fréquentiste). Pour rappel, un estimateur (en tant que fonction des données, lesquelles sont modélisées comme stochastiques) est une variable aléatoire. Exemple : l'estimateur MCO, $\widehat{\beta} \in (\mathbb{R}^{\dim(X)})^\Omega$.

Table des matières

1 Première partie – outils de probabilité et de statistique : les MCO

1.1 Estimation – Chapitre 1 : les fondamentaux de la régression linéaire

Régression linéaire théorique (ou projection) Sous des conditions de moment assez faibles (lesquelles sont supposées implicitement vérifiées, sauf mention explicite contraire, dès lors qu'on introduit une régression dans le cours, les TD, les examens et le reste de ce document) :

- moment d'ordre 2 fini pour la variable aléatoire réelle expliquée / de résultat Y : $\mathbb{E}[Y^2] < +\infty$;
- moment d'ordre 2 fini pour chacune des composantes du vecteur aléatoire colonne des variables explicatives / régresseurs / covariables X : $\mathbb{E}[\|X\|^2] = \mathbb{E}\left[\sum_{j=1}^{\dim(X)} X_j^2\right] < +\infty$ (où X_j désigne la j-ème composante de X) ;
- non-colinéarité parfaite des régresseurs : $\mathbb{E}[XX']$ inversible ;

on peut *toujours* définir la régression linéaire *théorique* de $Y \in \mathbb{R}^\Omega$ sur $X \in (\mathbb{R}^{\dim(X)})^\Omega$:

$$Y = X'\beta_0 + \varepsilon \text{ avec } \mathbb{E}[X\varepsilon] = 0,$$

où $\beta_0 := \mathbb{E}[XX']^{-1}\mathbb{E}[XY] \in \mathbb{R}^{\dim(X)}$ est un vecteur (colonne toujours) non stochastique (un paramètre), fonction de la distribution jointe $\mathrm{P}_{(Y,X)}$ du couple (Y, X), et $\varepsilon := Y - X'\beta_0 \in \mathbb{R}^\Omega$ est une variable aléatoire réelle.

On peut voir la régression linéaire théorique comme une projection orthogonale dans l'espace L^2 des variables aléatoires admettant une variance finie de la variable Y sur l'espace des fonctions *linéaires* de X. $\mathbb{E}[X\varepsilon] = 0$, équivalent à X orthogonal à ε, est la condition d'orthogonalité associée.

Estimation par MCO On peut estimer β_0 par la régression linéaire (empirique) correspondante :

$$\widehat{\beta} := \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right) = \widehat{\mathbb{E}}[XX']^{-1}\widehat{\mathbb{E}}[XY],$$

l'estimateur des Moindres Carrés Ordinaires (MCO) de Y sur X (rappel : $\widehat{\beta} \in (\mathbb{R}^{\dim(X)})^\Omega$ est un vecteur aléatoire car un estimateur, donc une statistique, c'est-à-dire une fonction mesurable des données, lesquelles sont modélisées comme stochastiques ; autrement dit, on peut bien calculer une réalisation de $\widehat{\beta}$ à partir d'un échantillon donné), sous les conditions de moments précédentes et avec des données indépendantes et identiquement distribuées $(Y_i, X_i)_{i=1,\dots,n} \overset{\text{i.i.d.}}{\sim} \mathrm{P}_{(Y,X)}$ (l'échantillonnage i.i.d est également toujours supposé sauf mention contraire dans le cours / TD / examens / ici),

- est bien défini avec une probabilité tendant vers 1 lorsque n tend vers l'infini, et
- converge en probabilité vers le coefficient théorique β_0, c'est-à-dire que $\widehat{\beta}$ est un estimateur consistant du paramètre β_0.

Par ailleurs, on a l'équivalent empirique de la condition d'orthogonalité de la régression théorique : $\widehat{\mathbb{E}}[X\widehat{\varepsilon}] = \overline{X\widehat{\varepsilon}} = n^{-1}\sum_{i=1}^n X_i\widehat{\varepsilon}_i = 0$ où $\widehat{\varepsilon}_i := Y_i - X_i'\widehat{\beta}$ est le résidu estimé de l'observation $i \in \{1, \dots, n\}$.

Référence clef du cours pour cette première partie : **Chapitre 1, Proposition 5** (définition de la projection linéaire théorique et estimation consistante par MCO).

Il n'y a pas (encore) de notion de causalité À ce stade, on ne parle *pas* de causalité ; on a juste une problématique de prédiction de Y à partir de fonctions linéaires de X.

Dans cette optique, le R^2 de la régression (Chapitre 1, slides 12 et 21) permet de quantifier la qualité de la prédiction de Y au moyen d'une fonction linéaire de X : $\widehat{Y} := X'\widehat{\beta} \xrightarrow[n\to+\infty]{P} X'\beta_0$.

Espérance conditionnelle et régression linéaire théorique En général, il n'y a aucune raison pour que $\mathbb{E}[Y \mid X] = X'\beta_0$, c'est-à-dire pour que l'espérance conditionnelle de Y sachant X soit linéaire.

Par contre, la régression linéaire théorique de Y sur X, qui donne la prédiction $X'\beta_0$, est la meilleure (au sens du risque quadratique / en norme L^2) approximation *linéaire* de $\mathbb{E}[Y \mid X]$ (Chapitre 1, Proposition 5, deuxième égalité du point 2) : $\beta_0 = \arg\min_{b \in \mathbb{R}^{\dim(X)}} \mathbb{E}\left[(\mathbb{E}[Y \mid X] - X'b)^2\right]$

Cas particulier important : la régression linéaire simple Lorsque X comporte uniquement une constante et un régresseur univarié / scalaire, c'est-à-dire une variable aléatoire réelle $D \in \mathbb{R}^\Omega$, soit formellement lorsque $X = (1, D)'$, l'estimateur MCO de la pente, noté $\widehat{\beta}_D$, c'est-à-dire l'estimateur du coefficient théorique associé à D (en notant $\widehat{\beta} = (\widehat{\alpha}, \widehat{\beta}_D)'$) dans la régression linéaire de Y sur D (et sur une constante, toujours implicitement présente sauf mention contraire) est égal à

$$\widehat{\beta}_D := \frac{\widehat{\mathbb{C}\mathrm{ov}}(Y, D)}{\widehat{\mathbb{V}}[D]} = \frac{(n-1)^{-1}\sum_{i=1}^n (Y_i - \widehat{\mathbb{E}}[Y])(D_i - \widehat{\mathbb{E}}[D])}{(n-1)^{-1}\sum_{i=1}^n (D_i - \widehat{\mathbb{E}}[D])^2},$$

et, sous les conditions de moment habituelles, il converge en probabilité vers les contreparties théoriques correspondantes :

$$\widehat{\beta}_D \xrightarrow[n\to+\infty]{P} \beta_D := \frac{\mathbb{C}\mathrm{ov}(Y, D)}{\mathbb{V}[D]},$$

en notant, dans ce cas d'une régression linéaire simple, $\beta_0 = (\alpha_0, \beta_D)'$ (Chapitre 1, slide 34).

Cas particulier du cas particulier : lorsque, de plus, D est binaire ($\mathrm{Support}(D) = \{0, 1\}$), alors

$$\beta_D = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0],$$

et on a l'expression empirique équivalente avec des $\widehat{\mathbb{E}}[\,\cdot\,]$ pour $\widehat{\beta}_D$ (Chapitre 1, slides 9 et 34).

Théorème de Frisch-Waugh et liens entre régressions linéaires simples et multiples Ce cas particulier donne une expression simple et assez compréhensible, intuitive

— (*Du côté des estimateurs aléatoires*) de l'estimateur MCO, $\widehat{\beta}_D \in \mathbb{R}^\Omega$, une variable aléatoire réelle,

— (*Le côté des paramètres non stochastiques*) et de sa limite en probabilité, le coefficient théorique $\beta_D \in \mathbb{R}$, un nombre réel non stochastique,

de la pente dans une régression linéaire simple (RLS) : c'est plus ou moins (à un facteur multiplicatif positif près) la corrélation (linéaire ou de Pearson) entre la variable expliquée Y et l'unique variable explicative ou régresseur D, puisque

$$\beta_D := \frac{\mathbb{C}\mathrm{ov}(Y, D)}{\mathbb{V}[D]} = \underbrace{\frac{\mathbb{C}\mathrm{ov}(Y, D)}{(\mathbb{V}[D]\mathbb{V}[Y])^{1/2}}}_{=\mathbb{C}\mathrm{orr}(Y,D)} \times \underbrace{\frac{\mathbb{V}[Y]^{1/2}}{\mathbb{V}[D]^{1/2}}}_{\text{constante} \geq 0},$$

et on a de même, avec les contreparties empiriques (symboliquement avec les chapeaux $\widehat{\cdot}$), pour $\widehat{\beta}_D$.

En comparaison, pour des régressions linéaires multiples (RLM), les expressions du paramètre théorique, β_0, et de son estimateur MCO, $\widehat{\beta}$, fournissent une expression explicite, mais moins intuitive. Le théorème de Frisch-Waugh (F.W) peut être interprété comme une façon de pallier ce problème et il est parfois qualifié de formule de "l'anatomie d'une régression (multiple)". En effet, il permet d'exprimer, pour un régresseur univarié d'intérêt, son coefficient théorique (ou son estimateur empirique MCO) comme le coefficient (sous-entendu, de la pente) d'une certaine régression linéaire *simple*.

Énoncé du théorème en version théorique (Chapitre 1, Proposition 6) – idem avec les contreparties empiriques dans la version "estimateurs MCO" (Chapitre 1, Proposition 3). Si l'on s'intéresse à un régresseur particulier $X_j \in \mathbb{R}^\Omega$ (une variable aléatoire réelle), pour $j \in \{1, \ldots, \dim(X)\}$, alors le coefficient (un nombre réel non stochastique) associé à X_j dans la régression linéaire théorique de Y sur X, qu'on note $\beta_{0j} \in \mathbb{R}$, la j-ème composante de β_0, vaut

$$(\beta_0)_{j\text{-ème comp.}} \overset{\text{noté}}{=} \beta_{0j} \overset{\text{F.W}}{=} \frac{\mathbb{C}\mathrm{ov}(Y, \xi)}{\mathbb{V}[\xi]} = \text{coefficient théorique (de la pente) dans la RLS de } Y \text{ sur } \xi,$$

où ξ est le terme d'erreur / résidus dans la régression linéaire théorique de X_j sur toutes les autres composantes de X, soit sur un vecteur aléatoire de dimension $\dim(X) - 1$ souvent noté X_{-j}.

ξ peut s'interpréter comme la variation "résiduelle", "nette" qui reste dans X_j en y retranchant ce qu'on peut expliquer par des fonctions linéaires de X_{-j}. β_{0j} s'interprète alors comme, à un facteur multiplicatif positif près, la corrélation entre la variable expliquée Y et le régresseur X_j net des variations de X_j expliquées par des fonctions linéaires des autres régresseurs X_{-j}.

De là également les formulations habituelles d'un effet marginal "toutes choses égales par ailleurs", "en laissant fixes les valeurs des autres régresseurs", pour l'interprétation des régressions multiples.

Un autre résultat important reliant RLS et RLM est la formule dite du "biais de variable omise" bien qu'à ce stade il faille voir qu'il n'y a pas de notion de biais : cette formule est uniquement une relation algébrique entre une régression "courte" et une régression "longue" :

— Chapitre 1, Proposition 4 pour la version empirique avec les estimateurs MCO,

— Chapitre 1, Proposition 7 pour la version théorique avec les coefficients de la régression linéaire théorique.

Au-delà des notations particulières du cours, il faut retenir "avec des mots" ce résultat pour mieux le retenir (comme tout résultat d'ailleurs de manière générale). En version synthétique (voir Quiz 1, Question 5 pour plus de détails), cela donne :

$$\textit{court} = \textit{long} + \textit{omises} \times \textit{coefficients des omises sur l'incluse}.$$

1.2 Inférence – Chapitre 2 : incertitude statistique dans les régressions linéaires

C'est bien d'avoir un estimateur consistant $\widehat{\beta}$ de β_0, mais cela reste un estimateur avec, pour un échantillon donné, uniquement une réalisation de cet estimateur (*estimator*) : une estimée (*estimate*). Que peut-on en déduire, qu'est-ce que cela permet d'apprendre sur le paramètre inconnu β_0 ? C'est la question de l'inférence étudiée au Chapitre 2 du cours.

Normalité asymptotique de l'estimateur MCO Le résultat fondamental de ce chapitre est, sous des conditions de moment qui restent faibles et l'échantillonnage i.i.d, la normalité asymptotique de l'estimateur MCO $\widehat{\beta}$ autour du coefficient théorique β_0 (**Chapitre 2, Théorème 1**) :

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0, \underbrace{\mathbb{E}[XX']^{-1}\,\mathbb{E}[\varepsilon^2 XX']\,\mathbb{E}[XX']^{-1}}_{=: V_a(\widehat{\beta}) \text{ la variance asymptotique de } \widehat{\beta}}).$$

Ce résultat est obtenu *sans supposer* l'*homoscédasticité* des résidus ε (relativement aux régresseurs X) : $\mathbb{E}[\varepsilon^2 XX'] = \mathbb{E}[\varepsilon^2]\,\mathbb{E}[XX']$ (Chapitre 2, Équation (Hom)). Tant mieux, car cette dernière condition est souvent restrictive en pratique en imposant que l'ampleur des erreurs de prédiction ne soit pas corrélée avec les régresseurs ; soit, dans l'idée, qu'il y ait autant de variabilité, de dispersion dans les Y quelle que soit la valeur de X. Un contre-exemple classique : Y = salaire, X = nombre d'années d'éducation, par exemple en France dans les années 2020.

Par défaut (et contrairement au défaut de Stata, attention), on utilise donc l'option `robust` dans la commande Stata `regress` afin de calculer des erreurs-types (utilisés dans les tests et les intervalles de confiance) robustes à l'hétéroscédasticité.

Précision de l'estimateur MCO En corollaires du Théorème 1, la première section du Chapitre 2 présente deux résultats donnant une expression plus compréhensible de la variance asymptotique de l'estimateur MCO, dans un cadre simplifié en supposant l'homoscédasticité.

(RLS) Chapitre 2, Équation (2) : dans une régression linéaire simple, $X = (1, D)'$, la variance asymptotique de l'estimateur MCO $\widehat{\beta}_D$ de la pente vaut

$$V_a(\widehat{\beta}_D) = \frac{\mathbb{E}[\varepsilon^2]}{\mathbb{V}[D]}.$$

(RLM) Chapitre 2, Proposition 1 : dans une régression linéaire multiple, la variance asymptotique de l'estimateur MCO $\widehat{\beta}_j$ d'un coefficient donné β_{0j} associé au régresseur univarié X_j vaut

$$V_a(\widehat{\beta}_j) = \frac{\mathbb{E}[\varepsilon^2]}{(1 - R^2_\infty)\mathbb{V}[X_j]},$$

où R^2_∞ est la limite en probabilité du R^2 de la régression de X_j sur les autres régresseurs X_{-j}.

Ces deux expressions sont valides sous l'hypothèse d'homoscédasticité. Plus généralement pour autant, on peut comprendre ces deux résultats comme une façon de rendre plus intuitif, plus compréhensible [1] l'expression de la (matrice de) variance(-covariance) asymptotique $V_a(\widehat{\beta})$.

Qualitativement, en termes de dépendances, ces résultats restent valides plus généralement, sans supposer l'homoscédasticité, et indiquent donc que la variance asymptotique d'un estimateur MCO

- croît (c-à-d, la précision de l'estimateur MCO décroît) lorsque la variance des résidus augmente ;
- croît (sa précision décroît) lorsque le régresseur considéré X_j est davantage corrélé avec les autres régresseurs X_{-j} ;
- décroît (sa précision croît) lorsque la variance du régresseur considéré augmente.

Tests et intervalles de confiance Sous réserve de disposer d'un estimateur consistant de la variance asymptotique $V_a(\widehat{\beta})$, ce qu'on a (moyennant une condition de moment supplémentaire) en remplaçant les espérances théoriques inconnues $\mathbb{E}[\cdot]$ par les moyennes empiriques $\widehat{\mathbb{E}}[\cdot]$ et le résidus théorique ε inobservé par les résidus estimés $\widehat{\varepsilon}$ (Chapitre 2, Proposition 2), la normalité asymptotique de $\widehat{\beta}$ permet de construire tests et intervalles de confiance pour (des composantes de) β_0 robustes à l'hétéroscédasticité et ayant les bonnes garanties asymptotiques :

- Chapitre 2, Proposition 3 : tests simples bilatéraux
- Chapitre 2, Proposition 4 : tests simples unilatéraux
- Chapitre 2, Proposition 5 : tests "multiples" ou "joints"
- Chapitre 2, Équations (8) et (9) : intervalles (pour des paramètres scalaires) et régions (pour des paramètres vectoriels) de confiance, ils ont des liens avec les tests (Chapitre 2, slide 33).

Interprétation d'une régression En pratique, il est attendu (et demandé) de savoir lire et commenter une sortie Stata de régression (voir les différents TD et, par exemple, Quiz 2, Question 11) en discutant notamment, pour un coefficient associé à un régresseur donné X_j, les trois éléments suivants :

1. Le signe du coefficient (interprétation qualitative) : est-il attendu ? surprenant ? quel sens de l'effet en termes de prédiction ? Plus tard, une fois introduite la causalité : au vu du signe, est-ce raisonnable de penser qu'il s'agit de l'estimation d'un effet causal (moyen) ?
2. Sa significativité statistique : c'est les réponses, ou la p-valeur pour les synthétiser, à la question du test bilatéral simple de nullité du coefficient théorique associé : $H_0 : \beta_{0j} = 0$ contre $H_1 : \beta_{0j} \neq 0$. Rappel : la *p-valeur* d'un test est le plus petit niveau pour lequel on rejette l'hypothèse nulle H_0 en faveur de l'hypothèse alternative H_1 (Chapitre 2, slides 21 et 25).

   Exemple. Si la p-valeur de ce test vaut $0.03 = 3\%$, alors :
   - $3\% > 1\%$: on dit que X_j n'est *pas* statistiquement significatif à 1% ;
   - $3\% < 5\%$: on dit que X_j est statistiquement significatif à 5% (et a fortiori à 10% ou à tout autre niveau plus élevé).

   Typiquement, si la p-valeur est plus petite que 1%, on dit que X_j est statistiquement significatif à tout niveau usuel (on a en tête 1%, 5%, 10%). Inversement, si la p-valeur est plus grande que 10%, on dit que X_j n'est statistiquement significatif à aucun niveau usuel [2].

1. De façon analogue au théorème de Frisch-Waugh (dans une perspective d'estimation par opposition ici à une problématique d'inférence) qui, pour une composante donnée, rend plus compréhensible l'expression du vecteur aléatoire $\widehat{\beta}$ (version empirique de F.W) ou du vecteur non stochastique β_0 (version théorique de F.W).

2. La p-valeur et la significativité statistique sont des notions importantes et utiles, mais ayant également leurs limites et sont critiquées (ou plutôt leurs mauvaises utilisations) – hors cadre du cours, voir néanmoins Quiz 2, Question 15.

3. La valeur du coefficient, la significativité pratique du coefficient (interprétation quantitative, et non seulement qualitative du signe). Pour cela :
   — Attention au modèle considéré :
     (*i*) level-level, log-level, level-log, log-log ?
     (voir Quiz 4, Question 5 pour les différentes interprétations)
     (*ii*) y a-t-il des régresseurs qui interviennent avec des puissances ou en interaction ?
     (voir Quiz 1, Question 12 sur les *effets marginaux*)
   — Attention aux unités des variables Y et X_j.
     Rappel : pour exprimer une variation *absolue* d'un pourcentage, on utilise les points de pourcentage, 1 p.p. $= 0.01$. Exemple : passer de 20% à 28% correspond à une augmentation *absolue* de 8 points de pourcentage et à une augmentation *relative* de 40% puisque $0.40 \times 0.20 = 0.08 = 28\% - 20\%$.
     On peut toujours (par construction des MCO et de la projection linéaire théorique) avoir une interprétation en termes de prédiction, *mais pas forcément en termes d'effet causal par contre !* (voir Partie 2 ci-dessous). D'où ce genre de formulations utiles pour commenter quantitativement la valeur d'un coefficient estimé : « Toutes choses égales par ailleurs, pour une augmentation de ... (en précisant bien les unités), on prédit une augmentation/diminution de ... (en précisant bien les unités) ».

Inférence sous des hypothèses plus fortes La Section 3 du Chapitre 2 ("Cas particuliers*") présente deux cas pour lesquels on peut s'attendre à améliorer l'inférence, au sens d'avoir une inférence plus précise : intervalles de confiance de même niveau (asymptotique), mais de longueur plus petite, toutefois au prix d'hypothèses plus fortes :

— homoscédasticité (souvent restrictif en pratique) ;

— résidus normaux / gaussiens et indépendants des régresseurs (souvent très restrictif en pratique).

Ce dernier cas permet également de construire tests et intervalles de confiance *exacts*, c'est-à-dire valides *non-asymptotiquement*, pour toute taille n d'échantillon contrairement aux garanties seulement asymptotiques. Pour plusieurs raisons (qu'on peut défendre, mais non discutées ici) l'approche dominante du cours pour l'inférence est une approche *asymptotique*, et robuste à l'hétéroscédasticité.

1.3 Utilisation pour la prédiction (dans un environnement stable) – Chapitre 3 : régressions linéaires et prédictions non causales

Prédictions non causales par opposition à prédictions causales (contrefactuelles) Le Chapitre 3 étudie l'utilisation des MCO pour faire des prédictions non causales *dans un environnement stable*, ce qui signifie formellement :

$$(Y_{n+1}, X_{n+1}) \overset{d}{=} (Y_i, X_i),\ \forall i \in \{1, \dots, n\} \overset{d}{=} (Y, X) \sim \mathrm{P}_{(Y,X)} \text{ (une instance générique de même loi)},$$

c'est-à-dire que les nouvelles variables, indexées par $n+1$, dont on cherche à prédire la composante Y à partir de X, et qui ne sont *pas* (*out-of-sample prediction*) dans l'échantillon initial, dont les observations sont indexées par $i = 1, \dots, n$, suivent la même loi que les observations de l'échantillon initial.

Ce cadre de prédiction non causale s'oppose à celui de prédictions dites "contrefactuelles" (Chapitre 0, Section 2). Ces prédictions contrefactuelles permettent de quantifier des relations causales puisque *la causalité sera définie en termes de variables contrefactuelles appelées variables potentielles.*

Au contraire, pour les prédictions non causales dans un environnement stable, on ne se pose pas de question de causalité. Autrement dit (voir Quiz 3, Question 10 pour plus de détails), on ne cherche pas à expliquer les mécanismes sous-jacents, véritables qui expliqueraient pourquoi X peut être utile pour expliquer Y ; *on cherche seulement à prédire Y au mieux (en termes de risque quadratique typiquement) à partir de fonctions (linéaires ici en utilisant les MCO) de X.*

Quelques thèmes principaux Vous reverrez et étudierez plus en profondeur ce problème dans vos cours d'apprentissage statistique (*machine learning*). Le Chapitre 3 peut être vu comme une introduction (avec déjà du contenu !) à ce problème dans le cas où on se restreint à des prédictions basées sur des modèles linéaires. En dépit de cette restriction, plusieurs notions ou idées importantes sont présentées et demeurent pertinentes et centrales dans un cadre plus général :

— Arbitrage dans l'erreur de prédiction out-of-sample (Chapitre 3, Théorème 1) entre

(i) un terme de "biais", décroissant en la complexité / richesse du modèle.

Question derrière : est-ce que, en laissant de côté l'incertitude statistique, le modèle est assez riche, suffisamment réaliste, pour que, en moyenne sur les échantillons possibles, les prédictions obtenues coïncident avec la prédiction optimale / oracle (en l'occurrence, avec l'espérance conditionnelle pour la perte quadratique) ?

(ii) un terme de "variance" , croissant en la complexité / richesse du modèle.

Question derrière : à quel point le modèle estimé obtenu et les prédictions qui en résultent varient, sont stables selon l'échantillon particulier tiré ?

Une autre formulation de cet arbitrage est l'opposition entre

(i) *under-fitting* (sous-interprétation) : le modèle n'est pas assez complexe ; le terme de biais domine.

(ii) *over-fitting* (surinterprétation) : le modèle est trop complexe, il apprend des variations aléatoires non pertinentes en "collant" à l'échantillon particulier de données utilisées et, dès lors, échoue à généraliser hors de l'échantillon ; le terme de variance domine.

Le Théorème 2 du Chapitre 3 donne ensuite une expression des termes de biais et de variance sous des hypothèses plus fortes (*toujours dans cette idée de simplifier afin d'obtenir des expressions plus compréhensibles*).

— Le principe de validation croisée pour choisir un modèle avec la séparation de l'échantillon initial en un échantillon d'apprentissage et un échantillon de validation / test.

— Le principe des régressions pénalisées et plus largement de la pénalisation :

— en "norme 0" : sélection des régresseurs et critères d'information ; cela pose des problèmes computationnels, car ce n'est pas un programme d'optimisation convexe ;

— pour régler cela, idée de relaxation convexe : un problème proche du problème initial et qui est convexe donc facilement calculable $\longrightarrow$ pénalisation en norme 1 (Lasso).

— pénalisation en norme 2 (régression Ridge).

2 Deuxième partie – définition des effets causaux et liens avec les régressions linéaires – Chapitre 4 : régressions linéaires et causalité, Sections 1 et 2

Début véritable de l'économétrie Cette partie correspond aux deux premières sections

— le cas d'une seule variable binaire

— le cas d'une seule variable non binaire

du Chapitre 4 : régressions linéaires et causalité. En un sens, le cours d'économétrie débute véritablement ici. Après les trois premiers chapitres (Partie 1) essentiellement statistiques, on rentre dans l'économétrie à proprement parler, qui se trouve à l'intersection des mathématiques appliquées (statistiques), mais aussi des sciences sociales (économie, etc.) ⟶ CLEF (DIFFICULTÉ ET INTÉRÊT) : *faire le lien entre une situation réelle (économique ou autre – langage ordinaire) et les propriétés (mathématiques – langage formelle) des variables aléatoires* (Chapitre 0, slide 4).

2.1 Formalisation de la notion de causalité via la notion de variables potentielles

Le début du Chapitre 4 formalise et définit la notion de causalité, plus précisément *l'effet causal* d'un traitement D sur une variable de résultat, "outcome", ou encore variable expliquée, Y.

Cette définition de l'effet causal est faite conjointement avec l'introduction des *variables potentielles de résultat* : les $Y(d)$ pour d une variable muette, une valeur possible, c'est-à-dire un nombre réel (dans le cas d'un traitement univarié, à ce stade), de la variable aléatoire réelle D : $d \in \text{Support}(D) \subset \mathbb{R}$.

Ces variables potentielles (de résultat) sont *contrefactuelles*. Elles représentent, pour chaque individu (les $Y(d)$ sont des $Y(d)_i$ pour chaque individu / unité / observation i), ce qui se serait passé pour la valeur de la variable de résultat pour un individu si celui-ci avait eu telle valeur, telle réalisation de la variable de traitement D (dans un univers parallèle si vous voulez en quelque sorte) [3].

$Y(d)_i$ est ainsi ce qu'*aurait eu* l'unité i pour la variable de résultat si elle avait eu $D_i = d$ pour sa variable de traitement.

L'effet causal individuel, c'est-à-dire spécifique à un individu, et ses variables potentielles de résultat sont *définis conjointement. Il faut accepter que ce soit la définition de la causalité dans ce cours (et dans l'économétrie contemporaine).*

Cas d'un traitement binaire Lorsque D est binaire, c'est-à-dire est une variable aléatoire réelle suivant une loi de Bernoulli, prenant la valeur 0 (non traité) ou 1 (traité), on définit *l'effet causal individuel de D sur Y* par

$$\Delta := Y(1) - Y(0) = \text{différence entre les variables potentielles de résultat.}$$

On observe seulement *la variable de résultat réalisée ou variable de résultat observée* définie par

$$Y := Y(D) = DY(1) + (1 - D)Y(0) = Y(0) + D[Y(1) - Y(0)] = Y(0) + D\Delta,$$

c'est-à-dire que $Y = Y(1)$ si $D = 1$ et $Y = Y(0)$ si $D = 0$.

Pour un même individu, il faut bien comprendre qu'on ne peut donc observer qu'une des deux variables potentielles, l'autre reste contrefactuelle. Sens de *contrefactuel* : c'est plus radical, plus problématique qu'une "donnée manquante", "non observée" ; ce n'est pas qu'on a mal réalisé la collecte des données et qu'on aurait pu observer cette variable, mais qu'on l'a manquée pour une raison ou une autre ; non, cette variable ne sera jamais accessible.

3. Une justification, ou plutôt une explication philosophique derrière cette construction (plus dans l'idée de faire comprendre le sens de ces variables potentielles) : « Quelque chose peut être le cas ou ne pas être le cas, et tout le reste demeurer inchangé. », *Tractatus logico-philosophicus*, Wittgenstein, proposition 1.21.

Autrement dit, une fois la réalisation (un ω est tiré) de la variable aléatoire D pour un individu, une et une seule de ses variables potentielles de résultat se réalise et est observée : $Y := Y(D)$, l'autre (dans le cas D binaire) ou toutes les autres (dans le cas général D non binaire) sont contrefactuelles.

Derrière un intérêt pour la causalité, il y a souvent un intérêt pour l'évaluation de politiques publiques ou de traitements : faut-il ou non mettre en œuvre un traitement ? Si Y représentait l'utilité d'un individu (nette du coût, en termes d'utilité, du traitement), un économiste voudrait le traiter, lui assigner le traitement D (c'est-à-dire avoir $D = 1$) si et seulement si Δ est positif. Idéalement, on voudrait donc connaître les effets causaux individuels Δ_i pour chaque individu i.

Sans plus d'hypothèses, le précédent paragraphe implique que c'est impossible : on ne peut espérer récupérer les Δ individuels. C'est pourquoi il faut se contenter d'effets causaux *agrégés*, qui sont formellement des fonctions de la distribution ou loi de probabilité P_Δ de Δ (entre autres dépendances selon les paramètres considérés). En effet, Δ étant spécifique à chaque individu et non observé, il est modélisé comme aléatoire, stochastique : mathématiquement, Δ est une variable aléatoire réelle (effets causaux *hétérogènes*).

Dans ce cours, on se concentre sur des effets causaux *moyens* (avec éventuellement des poids, moyenne pondérée, ou possiblement aussi une moyenne sur une sous-population seulement).

Dans le cas D binaire, les deux paramètres principaux d'intérêt sont :

$\delta := \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\Delta]$, l'effet causal moyen (sur toute la population d'intérêt),

$\delta^{\mathrm{T}} := \mathbb{E}[Y(1) - Y(0) \,|\, D = 1] = \mathbb{E}[\Delta \,|\, D = 1]$, l'effet causal moyen sur la sous-population des traités.

Cas d'un traitement univarié, pas forcément binaire Lorsque D n'est pas forcément binaire, mais que D a encore un sens quantitatif et un ordre, on suppose un effet causal *linéaire* de D sur Y (ou linéaire en des transformations connues de D et de Y, voir les modèles log-log, log-level, etc. ; par ailleurs, on peut aussi inclure des interactions ou des puissances pour relâcher cette hypothèse de linéarité) : il existe $d_0 \in \mathrm{Support}(D)$ et une unique variable aléatoire réelle $\Delta \in \mathbb{R}^\Omega$ tels que, pour tout $d \in \mathrm{Support}(D)$,

$$Y(d) - Y(d_0) = \Delta(d - d_0) \qquad \text{(Chapitre 4, Équation (Effets lin.))}. \tag{L}$$

Remarque : si c'est vrai pour un d_0, c'est vrai pour tout $d_0 \in \mathrm{Support}(D)$; d_0 joue juste ici le rôle d'une valeur de référence pour exprimer que l'effet causal individuel de D sur Y est linéaire.

Cette hypothèse de linéarité n'était pas requise dans le cas D binaire puisqu'alors, D ne pouvant prendre que deux valeurs, la linéarité était en quelque sorte automatiquement vérifiée : pour un individu donné, il y a un unique effet causal Δ, l'effet de passer de non traité ($D = 0$) à traité ($D = 1$).

Ce n'est plus le cas lorsque D n'est pas binaire. L'hypothèse (L) ("L" pour Linéaire) d'un effet causal linéaire permet alors de se ramener à un *unique effet causal de D sur Y* par individu : ce que cause sur la variable de résultat Y une augmentation de 1 du traitement D.

Quand D est une variable continue[4], Δ s'interprète comme un *effet causal marginal*[5] au sens où l'on a, pour tout $d \in \mathrm{Support}(D)$, $\dfrac{\partial Y(d)}{\partial d} = \Delta$.

Remarque : dans (L), Δ reste aléatoire, spécifique à chaque individu.

2.2 Représentation linéaire non causale (projection linéaire théorique) et représentation causale : coïncident-elles ? c'est-à-dire : y a-t-il un biais de sélection ?

Le résultat à retenir du cours s'il n'y en avait qu'un Moralement (pour résumer au sens de la morale à retenir d'une histoire, d'une fable ou de théorèmes et propositions mathématiques ici !), les **résultats fondamentaux du début du Chapitre 4**,

4. C'est-à-dire quand la variable aléatoire réelle D admet une densité par rapport à la mesure de Lebesgue.
5. Remarque : à ne pas confondre avec la notion d'effet marginal (non causal), qui peut être définie sans référence à la notion de causalité (voir Chapitre 1, slides 17, 18 et 36 et également Quiz 1, Question 12).

— **Chapitre 4, Proposition 1** pour le cas D binaire,

— **Chapitre 4, Proposition 2** pour le cas D non binaire,

disent qu'*on peut identifier et estimer de façon consistante par les MCO (la limite en probabilité de l'estimateur MCO de la pente est égale à l'effet causal moyen d'intérêt) un paramètre causal moyen (avec une moyenne éventuellement pondérée ou sur une sous-population uniquement) par une régression linéaire simple de Y sur D* SI *il n'y a pas de biais de sélection, c'est-à-dire si les variables potentielles de résultat $Y(d)$ ne sont pas corrélées avec la variable de traitement D*. Dans ce cas, on dit aussi que D est *exogène* relativement au modèle causal avec les variables potentielles de résultat.

Ce n'est donc PAS AUTOMATIQUE : « corrélation n'est pas causalité » comme on dit souvent, « No causation without manipulation » comme on dit parfois.

IL FAUT UNE HYPOTHÈSE ayant un sens concret (pas juste un sens mathématique, c'est là où l'économétrie rejoint les sciences sociales) pour que la régression linéaire de Y sur D estime de façon consistante un certain effet causal moyen de D sur Y : le traitement effectif D doit être généré, réalisé de telle sorte qu'il ne soit pas corrélé avec les variables potentielles de résultat $Y(d)$, $d \in \text{Support}(D)$.

Pour bien comprendre ce point et pourquoi on peut se poser cette question de la non-corrélation – ou plus largement de l'indépendance, mais, ici avec des régressions *linéaires*, la non-corrélation (*linéaire*) suffit intuitivement – entre D et $Y(d)$, pour $d \in \text{Support}(D)$, il faut bien se rendre compte que, conceptuellement, il n'y a aucun lien entre le traitement réalisé D et les variables potentielles de résultat $Y(d)$: ce sont deux objets distincts. Le résultat réalisé observé sera $Y := Y(D)$ mais, au niveau des variables potentielles, D et $Y(d)$ sont deux choses différentes et on se demande si le traitement D est déterminé d'une façon qui dépend ou non des variables potentielles $Y(d)$.

En règle générale, la détermination de D dépend des variables potentielles de résultat : il y a un biais de sélection et la limite en probabilité de l'estimateur MCO, le coefficient théorique β_D, mélange un effet causal et un effet de sélection : les individus traités qui ont $D = 1$ et les individus non traités qui ont $D = 0$ ne sont pas les mêmes (en termes de variables potentielles de résultats).

Voir également le paradoxe de Simpson (Quiz 4, Question 2).

Quelques précisions sur la formalisation mathématique de cette idée

Cas d'un traitement D binaire.
Chapitre 4, Proposition 1 :

$$\underbrace{\beta_D}_{\widehat{\beta}_D \xrightarrow[n\to+\infty]{P} \beta_D} = \delta^{\mathrm{T}} \iff \mathbb{C}\mathrm{ov}(D, Y(0)) = 0 \iff \underbrace{\mathbb{E}[Y(0) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 0]}_{=: B,\ \text{biais de sélection}} = 0.$$

Chapitre 4, slides 12 et 22 :

$$\underbrace{D \perp\!\!\!\perp (Y(0), Y(1))}_{D \text{ tiré aléatoirement (expérience)}} \implies \forall\, d \in \underbrace{\{0,1\}}_{\text{Support}(D)},\ \mathbb{C}\mathrm{ov}(D, Y(d)) = 0 \implies \underbrace{\beta_D}_{\widehat{\beta}_D \xrightarrow[n\to+\infty]{P} \beta_D} = \delta = \delta^{\mathrm{T}} = \delta^{W}.$$

Cas d'un traitement D non binaire. Chapitre 4, Proposition 2 :

$$\overbrace{(\mathrm{L})}^{\text{effets causaux linéaires}} \quad \text{et} \quad \overbrace{\forall d \in \text{Support}(D),\ \mathbb{C}\mathrm{ov}(D, Y(d)) = 0}^{\text{absence de biais de sélection}}$$
$$\implies \underbrace{\beta_D}_{\widehat{\beta}_D \xrightarrow[n\to+\infty]{P} \beta_D} = \delta^{W} := \mathbb{E}[W\Delta],\ \text{où } W := \frac{(D - \mathbb{E}[D])^2}{\mathbb{V}[D]}.$$

On peut lire ce dernier résultat de façon positive : sous ces deux conditions, une régression linéaire simple de Y sur D estime bien un paramètre causal, δ^W, un effet causal moyen *pondéré* avec un poids d'autant plus important pour un individu que son traitement D est différent, éloigné (distance

euclidienne au carré) du traitement moyen $\mathbb{E}[D]$. On peut aussi le lire de façon plus négative : même en l'absence de biais de sélection et avec des effets causaux linéaires, sans restreindre en rien l'hétérogénéité des effets causaux individuels (Δ est aléatoire et on n'impose pas de restriction sur la distribution P_Δ), alors au mieux une régression linéaire parvient à identifier un effet causal moyen, mais *pondéré*, avec des poids W qui ont l'avantage d'être bien positifs mais, pour autant, qui n'ont pas une interprétation claire quant à la décision de mettre en œuvre ou non le traitement étudié.

Effets causaux moyens et décision de mise en œuvre d'un traitement Quelques remarques préliminaires[6] sur ce dernier point. On néglige ici, pour simplifier, *l'incertitude statistique* : en réalité, on a au mieux une réalisation d'un estimateur consistant d'un paramètre causal et on peut faire de l'inférence (tests, intervalles de confiance) sur ce paramètre causal, mais on ne le connaît jamais avec certitude. On néglige également le coût du traitement ou, plutôt, on l'inclut dans Y qui représente l'utilité nette du coût de traitement d'un individu.

Considérons un planificateur égalitariste (il attribue la même pondération à chaque individu d'une population d'intérêt), bienveillant (ces pondérations sont positives), et utilitariste (il ne se préoccupe que du total ou, de façon équivalente, de la moyenne des utilités Y).

Il s'interroge sur la mise en œuvre sur cette population d'intérêt (celle de laquelle il a observé un échantillon représentatif) et, le cas échéant, de quelle manière, d'une certaine politique publique, le traitement D.

Il peut s'agir d'un "traitement" binaire (exemple : un programme de formation comme le JTPA du TD9), auquel cas il se demande s'il faut ou non traiter les individus, c'est-à-dire, qu'ils suivent le traitement ou programme en question. Plus largement, il peut aussi s'agir d'un traitement non binaire (exemple : le nombre d'années d'éducation) et le planificateur se demande alors s'il faut ajouter une (ou plusieurs) unités de ce traitement (suite de l'exemple : ajouter une année d'éducation supplémentaire en augmentant d'un an l'âge minimum d'arrêt des études) aux individus de la population considérée.

— $\delta > 0$ (en moyenne dans la population, l'effet causal individuel Δ est positif) est une condition nécessaire et suffisante (CNS) pour qu'il choisisse d'implémenter strictement le traitement au sens de forcer véritablement tous les individus de la population à le suivre effectivement dans le cas binaire : $D = 1$, ou, plus généralement, $D \to D + 1$ (augmenter de une unité, par exemple, l'intensité du traitement non binaire), puisque cela aura en moyenne, et donc aussi en total sur la population, un effet positif.

— $\Delta > 0$ presque sûrement (p.s) (l'effet causal individuel Δ est positif pour chaque individu) implique également cette mise en œuvre stricte, et même, dans ce cas, quelles que soient les pondérations (positives : le planificateur est bénévolent, il ne veut de mal à personne) des utilités intervenant dans le critère du planificateur.

— $\delta^{\mathrm{T}} > 0$ (en moyenne parmi les traités, l'effet causal individuel Δ est positif) est une CNS pour que le planificateur choisisse de donner libre accès au traitement, au sens de permettre à chacun de le suivre (dans le cas d'un traitement binaire, puisqu'on considère δ^{T}) si l'individu le décide, mais sans le forcer, puisque cela aura en moyenne sur les personnes traitées un effet positif[7].

— En revanche, $\delta^W > 0$ n'est *pas* une information suffisante à elle seule pour permettre au planificateur de décider d'implémenter ou non le traitement. On pourrait avoir $\delta^W > 0$ mais $\delta < 0$ et

6. Voir plus largement le champ de la théorie de la décision statistique (*statistical decision theory*), en prenant en compte l'incertitude statistique, ainsi que les problématiques de généralisation ("scalability") d'un traitement, d'une politique publique ; il y a de nombreuses questions liées qu'on néglige ici mais qui sont importantes en pratique : voir par exemple pour commencer les notions de *validité interne* par opposition à *validité externe* et les *effets d'équilibre général* ainsi que les *externalités* que vous connaissez déjà par votre formation d'économiste.

7. En information parfaite, si chaque agent peut connaître parfaitement (pas d'incertitude, qui demanderait aussi à introduire de l'aversion au risque, etc.) son Δ, et s'ils se conduisent de façon à maximiser leur utilité, les seuls individus qui décideront de suivre le traitement seront, de fait, ceux avec $\Delta > 0$ – (voir TD6, modèle 2 pour un tel modèle d'autosélection des individus dans le traitement ; le modèle 1 du TD6 est, au contraire, le cas d'une expérience aléatoire contrôlée).

$\delta^{\mathrm{T}} < 0$; $\Delta > 0$ p.s implique $\delta^W > 0$, mais la réciproque est bien sûr fausse. La condition $\delta^W > 0$ serait une CNS pour une mise en œuvre stricte du traitement si et seulement si le planificateur utilitariste utilisait les mêmes poids W intervenant dans l'effet causal moyen pondéré δ^W que dans son critère d'agrégation des utilités individuelles ; mais il n'y a aucune raison a priori de justifier un tel choix de pondérations [8].

Effets causaux hétérogènes ou homogènes On dit que les effets causaux (sous-entendu individuels) sont *hétérogènes* lorsque Δ est "aléatoire", c'est-à-dire formellement n'est pas une variable aléatoire réelle constante : $\mathbb{V}[\Delta] > 0$.

Dans le cas contraire, s'il existe un nombre réel δ_0 non stochastique tel que $\Delta = \delta_0$ p.s, autrement dit si P_Δ est une masse de Dirac en δ_0, on dit que les effets causaux sont *homogènes*.

Dans la plupart des applications, il s'agit d'une hypothèse forte, souvent peu réaliste. Elle a par contre l'intérêt de se ramener à un *unique* paramètre causal. On a ainsi, dans le cas D binaire (là où δ^{T} est naturellement pertinent) ou dans le cas général D non binaire,

$$\exists\, \delta_0 \in \mathbb{R} : \Delta = \delta_0 \text{ p.s} \implies \delta = \delta^{\mathrm{T}} = \delta^W = \text{toute moyenne pondérée ou sur une sous-population de } \Delta$$

puisque les effets causaux individuels sont tous égaux à δ_0.

L'hypothèse $\mathbb{E}[\Delta \mid D, G] = \delta_0$ des modèles (Mod. lin. 1) et (Mod. lin. 1) du Chapitre 4 autorise des effets causaux hétérogènes, mais en restreignant l'hétérogénéité : elle ne peut dépendre ni de D ni de G. L'hétérogénéité de l'effet causal peut être qualifiée d'idiosyncratique, car on autorise un effet causal aléatoire, spécifique à chaque individu, mais sans différence systématique due aux réalisations de leur traitement D et de leurs caractéristiques de contrôle G (voir Quiz 4, Questions 10 et 11 pour plus de détails sur cette hypothèse et, plus largement, sur les modèles linéaires 1 et 2 du Chapitre 4).

Dans la perspective du planificateur évoquée ci-dessus, cette restriction est importante, car elle permet, essentiellement, de se ramener à un *unique* effet causal *moyen* : δ_0 ; en particulier, $\mathbb{E}[\Delta] =: \delta = \delta_0$.

On peut relâcher, en partie, cette hypothèse en incluant des interactions dans le modèle (voir Chapitre 4, slide 29).

Le cas particulier (et une référence conceptuellement) des expériences aléatoires Il y a un cas particulier où l'on a bien absence de sélection / exogénéité de D : lorsque le traitement effectif réalisé D est "aléatoire" *au sens du langage ordinaire de ce terme* (D est toujours "aléatoire" au sens du langage mathématique formel d'être stochastique puisque D est une variable aléatoire), c'est-à-dire tiré, déterminé aléatoirement ; c'est le cas dans des expériences aléatoires contrôlées (*Randomized Controlled Trials* – RCT) où le traitement effectif D est bien "randomisé".

Par contre, dans des "expériences naturelles", des données administratives, des questionnaires, des observations dans la réalité sans toute la logistique, la manipulation d'une expérience aléatoire, la plupart du temps, notamment dès lors que les individus ont un certain contrôle sur D, choisissent d'une manière ou d'une autre (même partiellement) leur valeur réalisée de D, il y a des chances pour que D et les $\{Y(d)\}_{d \in \mathrm{Support}(D)}$, soient corrélés, c'est-à-dire pour que D soit *endogène*. Dans ce cas, il y a un biais de sélection : la régression linéaire simple de Y sur D n'estime *pas* de façon consistante un paramètre causal (ici défini comme une certaine moyenne des effets causaux individuels).

En pratique, VOUS POUVEZ ET DEVEZ VOUS POSER LA QUESTION SUIVANTE : *à quel point la façon dont est déterminée la variable de traitement effectif D dans les données qu'on observe est-elle proche ou éloignée d'une situation d'expérience aléatoire contrôlée dans laquelle le traitement D serait alloué aléatoirement, au hasard, tiré au sort ?*

À nouveau (*mais c'est crucial !*), en règle générale, dès lors que les individus ont une certaine liberté de choix pour D, on s'éloigne d'une situation d'expérience aléatoire contrôlée, dans laquelle au contraire

8. Comme dit plus haut, un égalitariste prendra des pondérations égales pour chaque individu ; d'autres philosophies politiques pourraient préférer d'autres pondérations, mais, a priori, elles n'ont aucune raison de coïncider avec les poids W, lesquels dépendent uniquement de la distance à l'intensité moyenne du traitement.

D s'impose aux individus. Hors expérience aléatoire avec "participation parfaite" (*perfect compliance*), c'est-à-dire lorsque le traitement effectif D est bien égal à l'assignation Z aléatoire (avec les notations du Chapitre 5), *cette hypothèse d'absence de sélection est souvent peu crédible*.

À partir de là, on peut voir **la troisième partie du cours** comme présentant deux façons d'essayer de résoudre ce problème et d'estimer un paramètre causal malgré la présence de biais de sélection.

3 Troisième partie – deux stratégies pour identifier un effet causal moyen malgré la présence d'un biais de sélection (inconditionnelle)

3.1 Ajouter les bonnes variables de contrôle – Chapitre 4, Sections 3 et 4

Identification et estimation consistante d'un effet causal moyen par une régression multiple Si $\mathbb{C}\text{ov}(D, Y(d)) = 0$ pour tout $d \in \text{Support}(D)$ est souvent peu crédible, sous réserve d'avoir de bonnes variables de contrôles G, il peut être plus crédible d'avoir *l'absence de sélection conditionnelle* : $\forall\, d \in \text{Support}(D), \mathbb{C}\text{ov}(D, Y(d) \,|\, G) = 0$.

Dans ce cas, en l'absence de sélection conditionnelle, toujours sous l'hypothèse d'effets causaux *linéaires* et sous réserve de *limiter l'hétérogénéité* des effets causaux individuels Δ en supposant $\mathbb{E}[\Delta \,|\, D, G] = \delta_0$, un nombre réel non stochastique, on peut bien estimer de façon consistante le paramètre causal d'intérêt qu'est l'effet causal moyen $\delta := \mathbb{E}[\Delta] = \delta_0$ à l'aide d'une régression linéaire *multiple* de Y sur D et G (**Chapitre 4, (Mod. lin. 1) et (Mod. lin. 2), Proposition 4**).

Biais de variable omise Au contraire, si l'on omet, dans les variables de contrôles, une variable

(a) qui affecte, influence, mathématiquement qui est corrélée avec la variable expliquée Y,

(b) ET (*il faut les deux*) qui est corrélée avec le traitement D,

alors on a un biais de variable omise (Chapitre 4, Proposition 5).

L'idée est donc d'avoir dans les contrôles G toutes les variables pouvant être corrélées à la fois à Y et à D, de façon à éviter un tel biais de variable omise et à obtenir l'absence de sélection conditionnelle. Autrement dit, avoir des variables G telles que, *conditionnellement à elles, on puisse considérer que la réalisation de D est comme si ("as-if") elle était déterminée au hasard, tirée aléatoirement* (derrière, il y a toujours la situation de référence conceptuellement d'une expérience aléatoire contrôlée).

Remarque : *attention*, cela ne veut pas dire qu'il faille chercher à mettre le plus de variables possible, n'importe quelle covariable dans G (voir Chapitre 4, slides 37 et 38 : biais de variables incluses).

Trouver de telles variables de contrôle G adéquates, d'une part les concevoir, avoir l'idée des bons contrôles et, d'autre part, être en capacité de les mesurer concrètement dans les données, est souvent difficile. D'où une autre approche par les méthodes de variables instrumentales.

3.2 Se servir d'un instrument valide – Chapitre 5 : variables instrumentales

Intuition Essentiellement, une variable instrumentale Z est une variable :

— qui affecte (formellement, qui est corrélée à) la variable de traitement D, y compris net d'éventuels contrôles G (*condition de pertinence*), et

— qui est exogène (*condition d'exogénéité*), au sens de non corrélée avec tous les facteurs inobservés, résumés dans le terme d'erreur η du cours (qui intervient dans les représentations linéaires causales avec les variables potentielles), jouant sur les variables potentielles de résultat $Y(d)$ (et donc jouant également sur la variable de résultat observée Y) :

  (i) on peut considérer que l'instrument Z est déterminé "comme si" il avait été tiré aléatoirement (éventuellement conditionnellement à des contrôles G), et

  (ii) Z affecte Y uniquement à travers D ; il n'y a pas d'effet, d'influence "propre" de Z sur Y (restriction d'exclusion).

Idée des résultats d'identification du Chapitre 5. Une variation de Z fait varier de façon exogène D qui, à son tour, est susceptible d'affecter et de faire varier Y. On obtient ainsi une variation exogène (au sens où l'on peut concevoir cette variation comme tirée au hasard – toujours dans cette référence au cadre conceptuel d'une expérience aléatoire contrôlée), ce qui, intuitivement, va permettre de retrouver un effet causal moyen de D sur Y *pour la sous-population de ceux qui sont affectés par l'instrument*, ceux dont le traitement D varie de façon exogène à la suite d'une variation de l'instrument Z.

Section 1 : expériences randomisées avec participation imparfaite (Z et D binaires, effets causaux hétérogènes, pas de contrôles G) On est dans le cadre d'instruments dont l'exogénéité (plus précisément, la partie (i)) est garantie par le fait qu'ils ont été tirés au sort :

— Z = l'indicatrice d'allocation initiale au traitement,

a priori *différente de*

— D = l'indicatrice de suivre effectivement le traitement (ce qui peut dépendre de choix de l'individu, le fait de s'inscrire, de suivre effectivement une formation par exemple, etc.).

Puisque D est déterminé (au moins en partie) par les agents, il est a priori endogène et il y a un biais de sélection. Par contre, l'assignation initiale Z est exogène, car tirée au sort (hypothèse d'indépendance, Équation (Indép.) du Chapitre 5). De plus, dans ce cadre, Z est bien corrélée (positivement) avec D (hypothèse $\mathbb{E}[D \mid Z = 1] > \mathbb{E}[D \mid Z = 0]$, Chapitre 5, Théorème 1).

Sous ces hypothèses, en supposant de plus la monotonie (Équation (Monot.) du Chapitre 5), alors on peut identifier au moyen d'une régression en deux étapes, Doubles Moindres Carrés (2MC), le paramètre causal δ^{C} : la moyenne des effets causaux individuels Δ *sur la sous-population des compliers*, les individus "conciliants" qui réagissent à l'instrument (**Chapitre 5, Théorème 1**).

On ne limite en rien à ce stade l'hétérogénéité des effets causaux individuels Δ. De ce fait, on ne peut identifier un effet causal moyen *que* sur un sous-ensemble de la population d'intérêt ; d'où le terme de LATE, "Local Average Treatment Effect", pour le paramètre causal δ^{C}.

Remarque : dans le cadre du Chapitre 4 avec D binaire, on pourrait aussi qualifier δ^{T} d'effet *local* puisqu'il s'agit de même d'une moyenne sur une sous-population particulière : les individus qui sont traités (lesquels sont a priori différents, en termes de variables potentielles de résultat, des individus non traités ; c'est le biais de sélection justement). Cette terminologie est toutefois réservée à δ^{C} plutôt.

Sections 2 et suivantes : généralisation et expériences "naturelles" (Z et D non-binaires, possiblement multivariés, possiblement des contrôles G, effets causaux homogènes) Par contre, à partir de la Section 2 du Chapitre 5, on simplifie en supposant des effets causaux *homogènes* (et *linéaires* toujours) :

$$Y(d) = \zeta_0 + \delta_0 d + \eta, \qquad \underbrace{\mathbb{E}[\eta] = 0}_{\text{sans perte de généralité car il y a une constante } \zeta_0} \qquad \text{(Chapitre 5, Modèle lin. 1)}$$

$$Y(d) = \zeta_0 + G'\gamma_0 + \delta_0 d + \eta, \quad \mathbb{E}[\eta] = \underbrace{\mathbb{E}[G\eta] = 0}_{\text{c'est-à-dire } G \textit{ exogène}} \qquad \text{(Chapitre 5, Modèle lin. 2)}$$

mais on ne suppose *pas* $\mathbb{E}[D\eta] = 0$: le traitement D peut être *endogène*.

L'hypothèse d'effets causaux homogènes implique : $\delta = \delta^{\mathrm{T}} = \delta^{W} = \delta^{\mathrm{C}} = \delta_0$.

Dans le Modèle 1 inconditionnellement, ou dans le Modèle 2 conditionnellement aux contrôles G, il faut noter que le seul élément stochastique de $Y(d)$, pour $d \in \mathrm{Support}(D)$, est le terme d'erreur η, qui agrège *l'hétérogénéité individuelle inobservée* affectant les variables potentielles de résultat $Y(d)$ (et donc aussi la variable de résultat observée Y).

Ainsi, pour toute variable aléatoire A, la covariance (conditionnellement à G) entre A et $Y(d)$ est égale à la covariance entre A et η, égale à $\mathbb{E}[A\eta]$ (ou $\mathbb{E}[A\eta \mid G]$) puisque η est centrée.

En particulier, par exemple inconditionnellement, sous l'hypothèse du Modèle lin. 1, on a :

$$\mathbb{C}\mathrm{ov}(D, Y(d)) = \mathbb{C}\mathrm{ov}(D, \eta) = \mathbb{E}[D\eta] \qquad \text{et} \qquad \mathbb{C}\mathrm{ov}(Z, Y(d)) = \mathbb{C}\mathrm{ov}(Z, \eta) = \mathbb{E}[Z\eta].$$

Dans ce cadre plus général des Sections 2 et suivantes, si les effets causaux sont *linéaires* et *homogènes* et si l'*instrument* Z de D est *valide*, alors on peut identifier et estimer de façon consistante l'effet causal homogène δ_0 par double moindres carrés :

- **Chapitre 5, Théorème 4** pour D univarié (un seul régresseur endogène) – identification,
- **Chapitre 5, Théorème 5** pour D multivarié (plusieurs régresseurs endogènes, dans ce cas il faut au moins autant d'instruments que de variables endogènes) – identification,
- **Chapitre 5, Théorème 6** – estimation et inférence : consistance et normalité asymptotique de l'estimateur $\widehat{\beta}_{\text{2MC}}$.

Instrument *valide* signifie qu'il satisfait les deux conditions : exogénéité et pertinence. Tandis qu'on ne peut pas tester l'exogénéité de l'instrument, on peut et il faut tester la condition de pertinence :

- Chapitre 5, Proposition 1 pour D univarié,
- Chapitre 5, Proposition 2 pour D multivarié,
- Chapitre 5, slide 39 pour le test en pratique.

3.3 Ouverture : d'autres stratégies d'identification existent – Économétrie 2 au second semestre et les cours d'économétrie de troisième année

Si l'on n'est pas dans un cadre d'expérience aléatoire (où l'allocation initiale Z au traitement est tirée au sort), de même qu'à la fin de la sous-section 3.1 de ce document (la stratégie d'ajout de bonnes variables de contrôle), on peut souvent douter de l'exogénéité de Z (même en contrôlant par G) : *il peut être difficile de trouver un instrument valide*.

Il existe encore d'autres façons de procéder, d'autres stratégies d'identification d'un effet causal. Nous en verrons certaines en Économétrie 2 au prochain semestre.

En particulier, on peut utiliser des données plus riches au sens où l'on peut suivre les mêmes individus dans le temps (*données de panel*) ou échantillonner une même population d'intérêt à plusieurs dates (*données de coupe répétées*). Voir Chapitre 0 : introduction, Section 3 : "Différents types de données" pour les définitions.

Remarque à ce propos : dans le cours d'Économétrie 1, on est toujours dans le cadre de *données de coupe*. De plus, on suppose toujours les échantillons i.i.d et, pour alléger les notations, on omet parfois les indices i pour désigner une variable ou un vecteur générique ayant la même loi que les données observées.

Pour les données de coupe se posent deux questions principales :

1. L'échantillon est-il bien "représentatif" de la population d'intérêt considérée ?
2. L'hypothèse d'indépendance, les vecteurs de variables correspondant aux différentes unités sont bien indépendants les uns des autres, est-elle crédible ?

Le cours d'Économétre 2 abordera le point 2 (notion de "clustering") et aussi en partie le point 1 (modèles de sélection). Le point 1 est crucial en pratique, même si le cours d'Économétrie 1 se concentre sur d'autres aspects également importants (voir Quiz 1, Question 13 et TD7, Questions 7 à 10 sur cette problématique de la représentativité d'un échantillon).

4 Compléments et conclusion

4.1 En vue des exercices théoriques notamment

Il est bien de connaître et de savoir utiliser (liste non exhaustive, quelques points fondamentaux) :

— la définition de la régression linéaire théorique,

— la loi des espérances itérées et sa généralisation (composition des projections),

— la définition des variables observées $Y := Y(D)$ (variable de résultat observée) ou $D := D(Z)$ (variable de traitement observée ; voir Chapitre 5, Section 1)

— la notion d'espérance conditionnelle et, plus largement, l'utilisation du conditionnellement.

Exemple Un exemple ci-dessous pour ces deux derniers points (la formulation précédente est peut-être un peu abstraite).

Dans le cadre d'un traitement D binaire (Chapitre 4, Section 1), le biais de sélection est défini par

$$B := \mathbb{E}[Y(0) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 0].$$

Question. Dans cette différence, l'une des deux espérances est contrefactuelle (on ne peut pas l'identifier) tandis que l'autre est identifiée dans les données, laquelle ?

Réponse. On a :

$$\begin{aligned} \mathbb{E}[Y \mid D = 0] &= \mathbb{E}[Y(D) \mid D = 0] \qquad \text{(par définition de } Y := Y(D)) \\ &= \mathbb{E}[Y(0) \mid D = 0] \qquad \text{(en utilisant le conditionnement).} \end{aligned}$$

Ainsi $\mathbb{E}[Y(0) \mid D = 0] = \mathbb{E}[Y \mid D = 0]$ est une fonction de la loi / distribution de probabilité $\mathrm{P}_{(Y,D)}$ des variables observées, et est donc une quantité identifiée dans les données.

Rappel : lorsqu'on parle d'identification, on se demande ce que l'on peut connaître *en faisant comme si* l'on connaissait la distribution des données, comme si l'on disposait d'un "échantillon infini".

Mieux, en plus d'être identifiée, on peut l'estimer de façon consistante par la moyenne empirique des Y_i sur le sous-échantillon des observations pour lesquelles $D_i = 0$ avec des données i.i.d (les conditions de moments, ici pour l'application de la loi des grands nombres, sont implicitement supposées vérifiées comme dans le cours parfois et comme toujours dans les TD ou les examens) :

$$\frac{\sum_{i=1}^n Y_i \mathbb{1}\{D_i = 0\}}{\sum_{i=1}^n \mathbb{1}\{D_i = 0\}} = \frac{\sum_{i \in \{1,\dots,n\}: D_i = 0} Y_i}{\mathrm{Card}(\{i \in \{1,\dots,n\} : D_i = 0\})} \xrightarrow[n \to +\infty]{P} \mathbb{E}[Y \mid D = 0] = \mathbb{E}[Y(0) \mid D = 0].$$

C'est juste un exemple d'un genre de raisonnement qui apparaît fréquemment dans les preuves des principaux théorèmes des Chapitres 4 et 5 et il est bien de savoir faire ce genre de raisonnement, notamment pour les exercices théoriques des examens et mi-parcours.

4.2 Un exemple

Puisqu'un bon exemple vaut souvent mieux que de longs discours… ci-dessous un exemple (à partir de données simulées) [9] pour illustrer le cœur du cours d'Économétrie 1 (voir également les exemples du cours et des TD, ainsi que les exemples du Quiz 4, Question 3).

La variable expliquée Y est la note finale obtenue à une matière de l'ENSAE. Le traitement D est le nombre d'heures de travail pour cette matière, moyenne hebdomadaire tout au long du semestre. On s'interroge sur l'existence d'un effet causal de D sur Y.

9. Vous pouvez me contacter pour le script `R` générant cet exemple si vous êtes intéressés.

FIGURE 1 – Plus travailler une matière donnerait en moyenne de moins bonnes notes... Pourquoi réviser ?!
Régression linéaire simple de Y sur D (et une constante) :

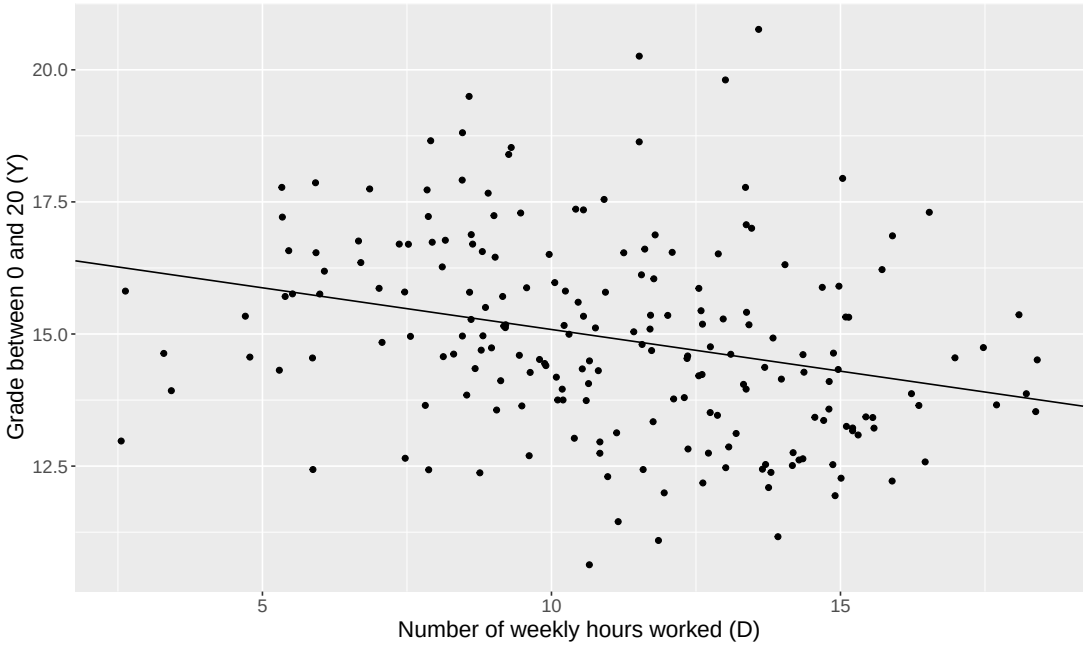
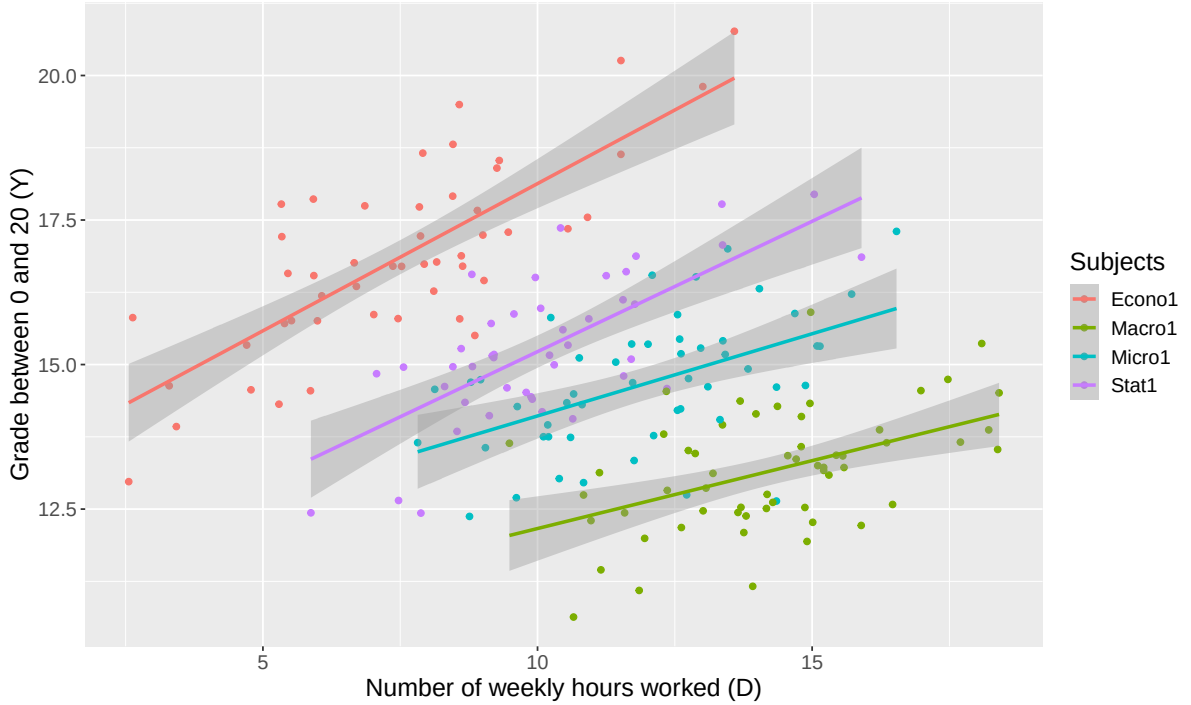


FIGURE 2 – Mais, peut-être que les étudiants ne choisissent pas au hasard, aléatoirement, le temps d'études consacré à une matière ! Il n'y a pas d'expérience aléatoire sur le temps de travail par matière ; autrement dit, les agents (ici les étudiants) choisissent eux-mêmes le "traitement". Peut-être même, hypothèse audacieuse, qu'ils travaillent davantage dans les matières les plus difficiles, où les notes ont, de toute manière, tendance à être plus basses. *Le paradoxe de Simpson* : paradoxe, car c'est évident une fois qu'on le dit, mais on l'oublie trop facilement avant de le voir, et, de plus, souvent, on n'observe pas dans les données ces contrôles pertinents.
Ici toutefois, on a comme contrôle G la matière, indiquée en couleur ci-dessous :

FIGURE 3 – La précédente régression linéaire simple souffre d'un biais de sélection : elle n'identifie pas un effet causal moyen de D sur Y. En effet, on a un biais de variable omise si on ne contrôle pas par la difficulté de chaque matière, qui est corrélée à la fois à la variable de traitement (le temps de travail D) et à la variable expliquée (la note Y). Ouf, avec ce contrôle, "le travail paye" : bonnes révisions ! Remarque : le graphique ci-dessous ne présente en fait pas une régression linéaire multiple avec le contrôle G, mais des régressions séparées selon les modalités de G (qui est discret ici), ce qui correspond davantage à la Proposition 3 du Chapitre 4 (slide 27) :



4.3 Pour résumer ce résumé (4 images et 160 mots environ)

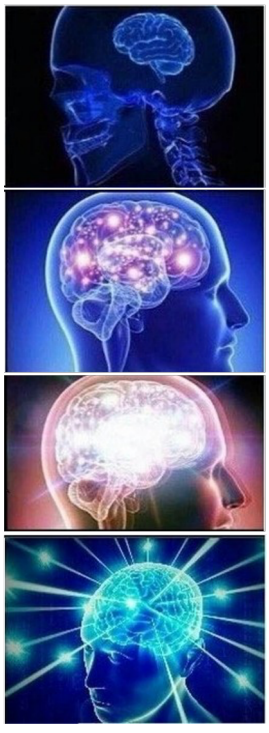


FIGURE 4 – Un "mème-sumé" du cours [10] :

La régression linéaire théorique peut *toujours* être définie sous de simples conditions de moment : un "modèle linéaire" au sens d'écrire (P) $Y = X'\beta_0 + \varepsilon$ avec $\mathbb{E}[X\varepsilon] = 0$ est en cela tautologique.

Mais cette représentation linéaire (P), "simple projection", *ne coïncide pas en général* avec la représentation causale faisant intervenir les variables potentielles de résultat $Y(d)$ et les paramètres causaux, qui sont *définis conjointement : effet causal* :=: *différences des* $Y(d)$.

Pour cela et pour estimer un effet, supposé linéaire, causal moyen (pondéré ou sur une sous-population) par une régr. lin. simple (MCO), *il ne doit pas y avoir de biais de sélection.* Exemple et référence conceptuelle : expériences où le traitement D est tiré aléatoirement.

La plupart du temps, cette absence de biais de sélection est peu crédible. Mais on peut néanmoins identifier des effets causaux (i) en ajoutant des variables de contrôle G adéquates (absence de sélection conditionnelle) ou (ii) en utilisant des variables instrumentales Z.

10. De ce que je comprends actuellement, il me semble qu'il existe un cinquième stade, évoqué partiellement dans ce résumé d'ailleurs, mais hors cours, et il en existe peut-être un sixième et d'autres encore : bonnes réflexions !