Économétrie des données de panel

M2 Économie de la Santé et Économie Appliquée

Université Paris Cité

Carla Morvan

2026-2027

Part II

Les modèles de panel usuels

Les modèles de panel usuels

Plan

Objectifs du chapitre

À l'issue de cette séance, vous serez capables de :

- Expliquer les limites du *pooled OLS* sur données de panel
- Estimer et interpréter un modèle à **effets fixes**
- Estimer et interpréter un modèle à **effets aléatoires**
- Utiliser le test de **Hausman** pour arbitrer entre les deux
- Estimer ces trois modèles sous R (`plm`, `fixest`) et comparer les résultats

Rappel : le modèle linéaire par MCO

On cherche à expliquer une variable y en fonction d'une (ou plusieurs) variable(s) x :

Le modèle

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- y_i : variable à expliquer
- x_i : variable explicative
- β_0, β_1 : paramètres à estimer
- ε_i : terme d'erreur (tout ce qui explique y_i en dehors de x_i)

Les MCO estiment β_0, β_1 en minimisant la somme des carrés des résidus :

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Rappel : les hypothèses des MCO

Pour que $\widehat{\beta}$ soit un bon estimateur (sans biais, efficace), plusieurs hypothèses doivent être vérifiées :

- **Linéarité** : le modèle est linéaire en les paramètres
- **Exogénéité** : $\mathbb{E}[\varepsilon_i \mid x_i] = 0 \rightarrow$ l'erreur n'est pas corrélée avec x_i
- **Homoscédasticité** : $Var(\varepsilon_i \mid x_i)$ constante
- **Pas d'autocorrélation** : les erreurs de deux observations différentes ne sont pas corrélées entre elles
- **Pas de colinéarité parfaite** entre les variables explicatives

L'hypothèse clé : Si l'**exogénéité** est violée, $\widehat{\beta}$ est **biaisé**
$\rightarrow$ ce n'est pas juste un problème de précision, c'est un problème d'estimation du bon effet.

Et avec des données de panel ?

Ce modèle simple ($y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$) suppose que chaque observation i est **indépendante** des autres.

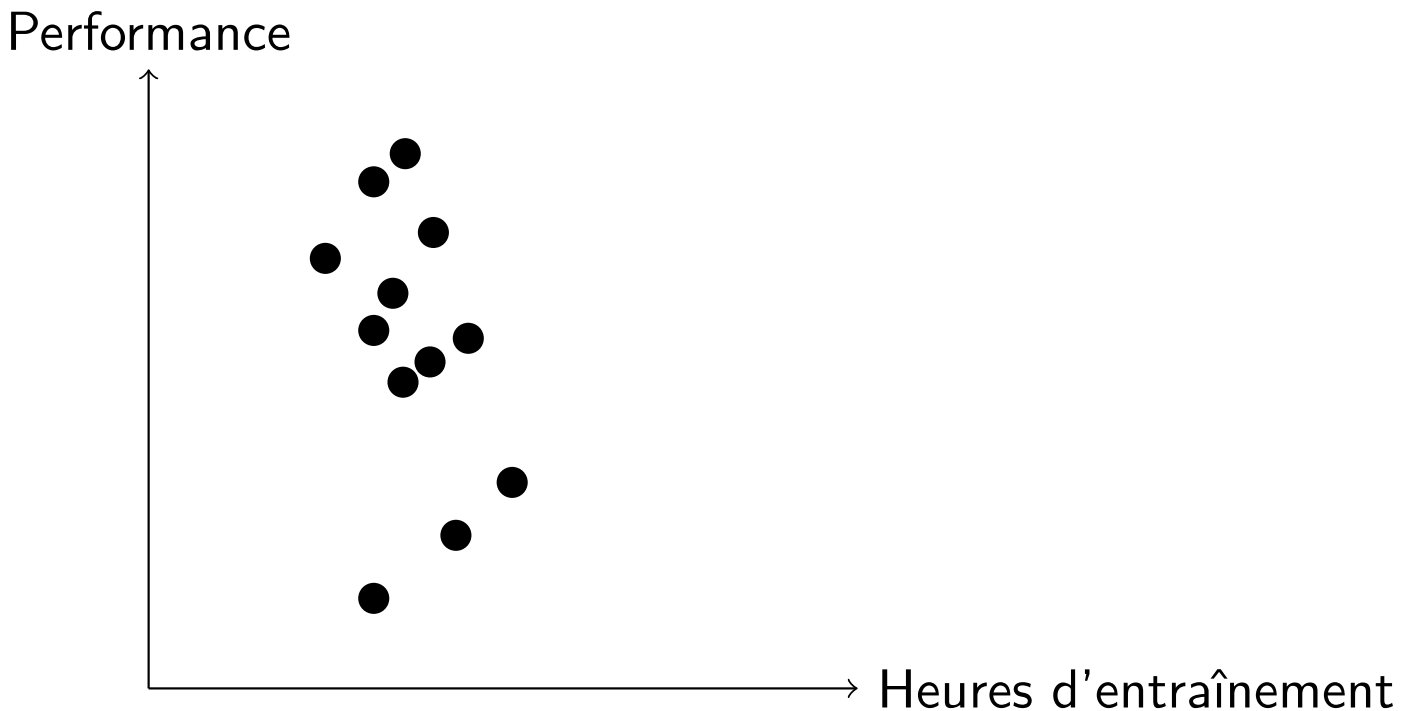
→ Avec un panel, on a **plusieurs observations pour un même individu** ($y_{i1}, y_{i2}, \ldots, y_{iT}$). Rien dans ce modèle ne tient compte de cette structure, il traite chaque paire (i, t) comme si elle venait d'un individu différent.

→ Or une partie de ε_{it} peut correspondre à des caractéristiques **propres à l'individu** i, qui ne changent pas dans le temps, et qui risquent d'être corrélées avec x_{it}.

⇒ **violation de l'hypothèse d'exogénéité** : $\mathbb{E}[\varepsilon_{it} \mid x_{it}] \neq 0$. Il faut adapter le modèle à la structure du panel.

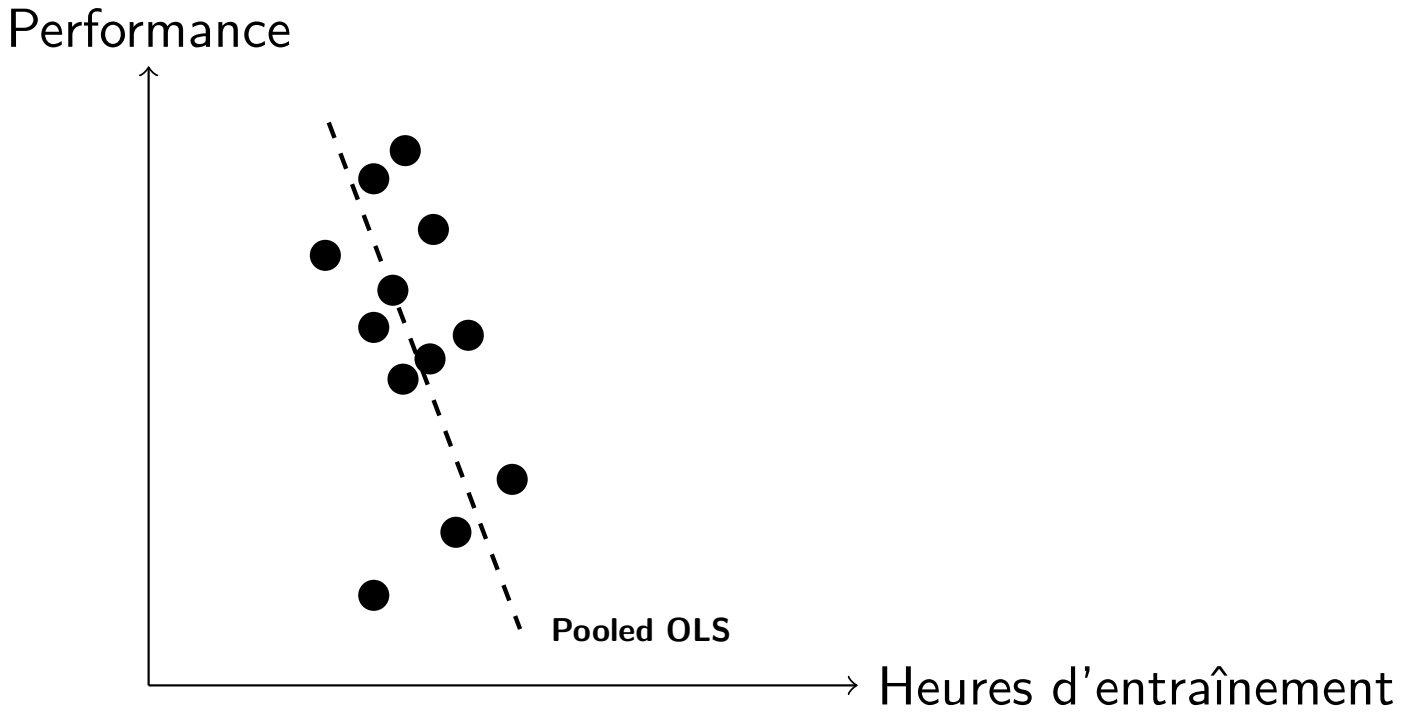
Le paradoxe de Simpson

On étudie l'effet des heures d'entraînement sur la performance sportive : intuitivement, **plus on s'entraîne, meilleure est la performance**.



Le paradoxe de Simpson

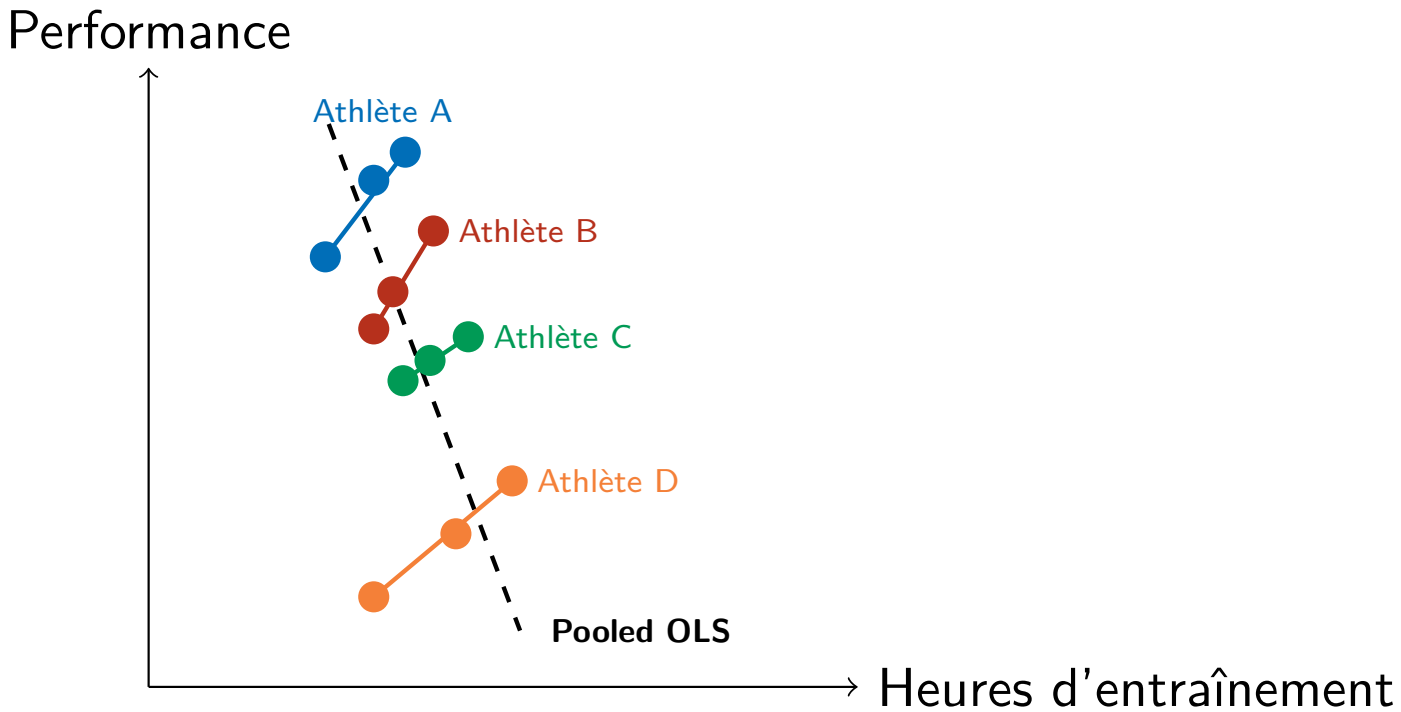
On étudie l'effet des heures d'entraînement sur la performance sportive : intuitivement, **plus on s'entraîne, meilleure est la performance**.



On observe une pente **négative** : plus d'entraînement → moins bonne performance ? Contre-intuitif...

Le paradoxe de Simpson

On étudie l'effet des heures d'entraînement sur la performance sportive : intuitivement, **plus on s'entraîne, meilleure est la performance**.



En réalité, **chaque athlète** progresse bien avec l'entraînement.
Le pooled OLS confond **niveau de base** (between) et **progression individuelle** (within).

Plan

Le modèle *pooled*

On adapte le modèle à la notation panel, mais sans encore exploiter sa structure :

Le modèle pooled

$$y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}$$

On estime ce modèle par MCO, en empilant toutes les observations (i, t) comme si elles étaient indépendantes : **un seul β_0, un seul β_1 pour tout le panel**.

→ Ce qu'on a vu avec le paradoxe de Simpson
Si ε_{it} contient une composante propre à l'individu i, corrélée avec x_{it}, alors $\mathbb{E}[\varepsilon_{it} \mid x_{it}] \neq 0$: le pooled OLS est **biaisé**, parfois même de signe opposé à l'effet réel.

Décomposer l'erreur

Le paradoxe de Simpson vient d'un problème précis : ε_{it} contient en réalité **deux choses différentes**. On les sépare explicitement :

Décomposition de l'erreur

$$\varepsilon_{it} = \alpha_i + u_{it}$$

- α_i **effet individuel** : tout ce qui caractérise l'individu i et **ne varie pas dans le temps** (niveau de base, culture...)
- u_{it} erreur **idiosyncratique** : ce qui varie dans le temps, propre à chaque observation (i, t) (exogène)

Le problème du pooled OLS
Si α_i est corrélé avec x_{it} (ex. : les athlètes moins doués s'entraînent plus), l'estimateur pooled est biaisé. Il traite α_i comme s'il faisait partie d'un bruit aléatoire non corrélé.

Le modèle à effets fixes individuels

Le modèle FE

$$y_{it} = \beta x_{it} + \alpha_i + u_{it}$$

Dans un modèle à effets fixes, α_i n'est plus dans l'erreur, il devient une **composante du modèle** : une constante propre à chaque individu.

Maintenant, il n'est plus nécessaire que $\mathbb{E}[\alpha_i \mid x_{it}] = 0$, α_i **peut** être corrélé avec x_{it}.

Concrètement : chaque individu i a sa **propre ordonnée à l'origine** (α_i), mais tous les individus partagent la **même pente** β, seul le niveau de y_{it} change d'un individu à l'autre, pas l'effet marginal de x_{it}.

Les effets fixes temporels

L'effet fixe individuel α_i capture ce qui est propre à chaque individu et invariant dans le temps. Mais symétriquement, il peut exister des **chocs communs à tous les individus, à une période donnée**.

> Exemple : une crise sanitaire mondiale (ex. Covid-19) affecte **tous les pays la même année**, un choc temporel, pas individuel.

→ On ajoute donc un second effet fixe, δ_t, propre à chaque période t (et commun à tous les individus).

Le modèle à effets fixes bidirectionnels (TWFE)

Le modèle TWFE

$$y_{it} = \beta x_{it} + \alpha_i + \delta_t + u_{it}$$

- α_i effet fixe **individuel** : neutralise l'hétérogénéité propre à chaque individu, invariante dans le temps
- δ_t effet fixe **temporel** : neutralise les chocs communs à tous les individus, à une période donnée

On contrôle ainsi **simultanément** pour tout ce qui varie entre individus (mais pas dans le temps) **et** tout ce qui varie dans le temps (mais pas entre individus).

Ce qui reste dans β : l'effet de x_{it} sur y_{it}, net de ces deux sources d'hétérogénéité inobservée.
→ c'est en général la spécification la plus robuste et la plus utilisée.

Une autre façon de voir l'hétérogénéité individuelle

Avec les effets fixes, α_i est un **paramètre à estimer**, propre à chaque individu et potentiellement corrélé avec x_{it}.

→ Une autre approche : traiter α_i comme une **variable aléatoire**, tirée d'une distribution commune à tous les individus, plutôt que comme un paramètre fixe et individuel.

L'idée derrière le modèle : les individus ne sont pas fondamentalement différents les uns des autres, **ils sont juste des tirages aléatoires issus d'une même population**. Leur hétérogénéité est due au hasard, pas à des caractéristiques systématiquement corrélées avec x_{it}.

Le modèle à effets aléatoires

Le modèle EA

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + u_{it}$$

Même écriture qu'avant, mais l'hypothèse sur α_i change radicalement :

Hypothèse : α_i suit une distribution aléatoire, **indépendante** de x_{it}

$$\alpha_i \sim iid(0, \sigma_\alpha^2), \qquad \mathbb{E}[\alpha_i \mid x_{it}] = 0$$

Autrement dit : **connaître x_{it} ne donne aucune information sur la valeur probable de α_i.**

Attention : en sciences sociales, c'est une hypothèse **très forte**. Les caractéristiques individuelles (richesse, institutions, talent...) sont rarement de purs tirages aléatoires indépendants de x_{it}.

Les modèles de panel : récapitulatif

Modèle	**Équation**	**Hypothèse clé**
Pooled	$y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}$	$\mathbb{E}[\varepsilon_{it} \mid x_{it}] = 0$
EF	$y_{it} = \beta x_{it} + \alpha_i + u_{it}$	α_i **peut** être corrélé avec x_{it}
TWFE	$y_{it} = \beta x_{it} + \alpha_i + \delta_t + u_{it}$	α_i et δ_t **peuvent** être corrélés avec x_{it}
EA	$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + u_{it}$	$\mathbb{E}[\alpha_i \mid x_{it}] = 0$

Les modèles de panel : récapitulatif

Modèle	**Équation**	**Hypothèse clé**
Pooled	$y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}$	$\mathbb{E}[\varepsilon_{it} \mid x_{it}] = 0$
EF	$y_{it} = \beta x_{it} + \alpha_i + u_{it}$	α_i **peut** être corrélé avec x_{it}
TWFE	$y_{it} = \beta x_{it} + \alpha_i + \delta_t + u_{it}$	α_i et δ_t **peuvent** être corrélés avec x_{it}
EA	$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + u_{it}$	$\mathbb{E}[\alpha_i \mid x_{it}] = 0$

→ **Comment choisir entre ces modèles ?**

1. La structure des données : ai-je vraiment plusieurs observations par individu, avec une hétérogénéité potentielle entre eux ? Si oui, le *pooling* est risqué
2. La question posée : est-ce que je soupçonne une hétérogénéité inobservée **au niveau individu** (α_i), **au niveau temporel** (δ_t), ou les deux et est-elle susceptible d'être corrélée avec x_{it} ?

Plan

Modèle vs estimateur

Avant d'estimer quoi que ce soit, il faut bien distinguer deux notions qu'on mélange souvent :

Le modèle

Hypothèses sur la **structure des données** (sur les potentielles corrélations)

L'estimateur

La **méthode de calcul** utilisée pour obtenir $\widehat{\beta}$ à partir des données (exemple : MCO).

- Un même modèle peut être estimé de plusieurs façons
- Les estimateurs diffèrent selon qu'ils exploitent la variation **between** ou **within** des données
- La validité d'un estimateur dépendent de **si le modèle est le bon**.

Deux critères pour juger un estimateur

Rappel: la validité d'un estimateur dépend du modèle. On juge sa consistance et son efficience **sous l'hypothèse qu'un modèle donné est vrai**. Ce ne sont pas des propriétés absolues de l'estimateur seul.

Consistance

Si le modèle est vrai, $\widehat{\beta}$ converge vers β quand $N \to \infty$.

Est-ce que je vise la bonne cible ?

Efficience

Parmi les estimateurs consistants, lequel a la plus petite variance.

Suis-je précis autour de cette cible ?

Exemple : Si le modèle EA est le vrai, l'estimateur *within* reste consistant, mais moins efficient que l'estimateur GLS. On choisira donc GLS.

L'estimateur pooled OLS

L'estimateur pooled OLS exploite **à la fois** la variation *between* et la variation *within* pour estimer les paramètres.
On l'obtient en empilant les données sur i et t en une seule régression de NT observations, estimée par MCO :

$$y_{it} = \alpha + x'_{it}\beta + \varepsilon_{it}$$

Si le vrai modèle est le modèle pooled **et** que les régresseurs ne sont pas corrélés avec le terme d'erreur, l'estimateur pooled OLS est **consistant**.

Si le vrai modèle est à effets fixes, l'estimateur pooled OLS est **inconsistant**.

→ Il faut néanmoins des **erreurs standards corrigées pour la structure de panel** (*panel-corrected standard errors*). [chapitre 3]

L'estimateur *between*

Une autre façon d'exploiter le panel : au lieu de garder toutes les observations (i,t), on **écrase la dimension temporelle** en ne gardant que la moyenne individuelle de chaque variable.

La transformation *between*

$$\overline{y}_i = \beta_0 + \overline{x}_i'\beta + \overline{\varepsilon}_i$$

On régresse en MCO les moyennes individuelles $\overline{y}_i$ sur $\overline{x}_i$. Une seule observation par individu, N observations au total (au lieu de NT).

Consistant **seulement si** α_i n'est pas corrélé avec x_{it}.
Cet estimateur est rarement utilisé en pratique, car les estimateurs *pooled* et *effets aléatoires* sont plus efficients.

L'estimateur *within*

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + \varepsilon_{it}$$

Symétrique de l'estimateur *between* : au lieu d'écraser la dimension temporelle, on **élimine la dimension individuelle**, en soustrayant à chaque observation sa moyenne individuelle.

La transformation *within*

$$y_{it} - \overline{y}_i = \beta_1 (x_{it} - \overline{x}_i) + \underbrace{(\alpha_i - \overline{\alpha}_i)}_{=0} + (\varepsilon_{it} - \overline{\varepsilon}_i)$$

α_i disparaît car il est **constant dans le temps** : $\overline{\alpha_i} = \alpha_i$.

Consistant **que α_i soit corrélé ou non** avec x_{it}, c'est précisément ce qui en fait l'estimateur des **effets fixes**.
Contrepartie : on perd toute la variation *between*

et dans le cas TWFE : qu'est-ce qu'on retire vraiment ?

Question : dans un modèle à effets fixes bidirectionnels, si on retire à la fois α_i et δ_t, est-ce qu'il ne reste plus rien ?

La transformation *two-way within*

$$y_{it} - \overline{y}_i - \overline{y}_t + \overline{\overline{y}} = \beta_1 (x_{it} - \overline{x}_i - \overline{x}_t + \overline{\overline{x}}) + \dots$$

La variation *between* était déjà retirée par $\overline{y}_i$ (comme dans le within simple). Soustraire $\overline{y}_t$ ne retire **pas** une nouvelle variation, elle retire seulement la **partie commune à tous les individus, à un instant** t (un choc partagé), qui restait cachée à l'intérieur de la variation within.

→ **Ce qu'il reste** : la variation **propre à l'individu** i**, à la période** t, purgée à la fois de son niveau individuel moyen **et** des chocs communs à cette période.

Least Squares Dummy Variable (LSDV) : une alternative équivalente

Plutôt que de transformer les données, on peut ajouter directement $N-1$ variables indicatrices (une par individu) :

Le modèle LSDV

$$y_{it} = \beta_1 x_{it} + \sum_{i=2}^{N} \gamma_i D_i + \varepsilon_{it}$$

LSDV donne **exactement le même** $\widehat{\beta}_1$ que l'estimateur *within*, ce sont deux méthodes de calcul pour le même résultat.

En pratique : rarement utilisé tel quel si N est grand (des centaines de dummies), mais utile conceptuellement pour "voir" les α_i comme de vrais coefficients estimés.

Une limite de l'estimateur *within*

Les variables **invariantes dans le temps** disparaissent du modèle, et leur coefficient n'est **pas identifié**.

Exemple : une variable indicatrice *homme/femme* (1/0), constante pour un individu donné sur toute la période. Sa valeur moins sa moyenne vaut **toujours 0**. Elle est éliminée par la transformation *within*, exactement comme α_i.

On ne peut donc pas faire l'analyse de l'effet du genre ou de l'origine ethnique sur le salaire avec un tel modèle, ou tout autre caractéristique individuelle ou invariante dans le temps.

Si on s'intéresse à l'effet d'une variable invariante dans le temps, il faut se tourner vers d'autres modèles : **pooled OLS** ou **estimateur between**.

L'estimateur en différences premières

Une autre façon d'éliminer α_i, Plutôt que de soustraire la moyenne individuelle (*within*), on peut soustraire **la période précédente** :

Le modèle de départ (deux périodes consécutives)

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + \varepsilon_{it}$$

$$y_{i,t-1} = \beta_0 + \beta_1 x_{i,t-1} + \alpha_i + \varepsilon_{i,t-1}$$

On soustrait les deux équations :

$$\Delta y_{it} = \beta_1 \Delta x_{it} + \Delta \varepsilon_{it}$$

α_i disparaît pour la même raison que dans le *within* : constant dans le temps, il s'annule dès qu'on prend une différence.

Estimateur consistant pour les meme raisons que within et meme difficulté pour les variable invariante dans le temps.

Différences premières vs *within*

1. Utilisation des données
   - Avec $T = 2$, les deux estimateurs sont **strictement identiques**
   - Avec $T > 2$, ils **diffèrent**
     - le *within* utilise toute l'information disponible
     - les différences premières ne comparent que deux périodes consécutives à la fois.
2. structure de l'erreur
   - ε_{it} iid (pas d'autocorrélation)
     - Le *within* est efficace.
     - Les 1D sont moins bonnes : $\Delta\varepsilon_{it}$ crée une autocorrélation **artificielle** entre soustractions consécutives.
   - ε_{it} suit une marche aléatoire (autocorrélé)
     - $\varepsilon_{it} = \varepsilon_{i,t-1} + \eta_{it}$: différencier **élimine** cette accumulation et redonne des erreurs bien comportées.

En pratique : le cas des erreurs peu autocorrélées domine en **micro-panel** (*within*). Alors qu'en **macroéconométrie** les séries sont souvent non-stationnaires (1D).

L'estimateur des effets aléatoires (GLS)

On a vu deux estimateurs extrêmes : **between** (ne garde que la variation entre individus) et **within** (ne garde que la variation dans le temps).

L'estimateur des effets aléatoires est une **combinaison pondérée** des 2 :

L'estimateur Moindres Carrés Généralisés MCG (*feasible GLS*)

$$y_{it} - \theta \overline{y}_i = \beta_0(1-\theta) + \beta_1(x_{it} - \theta \overline{x}_i) + [\varepsilon_{it} - \theta \overline{\varepsilon}_i]$$

où $\theta \in [0,1]$ dépend du rapport entre variance individuelle (σ^2_α) et variance idiosyncratique (σ^2_u) de l'erreur $\varepsilon_{i,t} = \alpha_i + u_{i,t}$: t.q :

$$\theta = 1 - \sqrt{\frac{\sigma^2_u}{\sigma^2_u + T\sigma^2_\alpha}}$$

$\theta = 1 \rightarrow$ estimateur **within**. $\theta = 0 \rightarrow$ estimateur **pooled**.
Entre les deux, le GLS pondère selon la fiabilité relative de chaque source de variation.

Pourquoi le MCG est-il plus efficace ?

Si α_i n'est vraiment pas corrélé avec x_{it} (hypothèse des effets aléatoires), alors la variation *between* contient de l'information **utile**, pas seulement du bruit à éliminer.

L'estimateur *within* jette cette information par sécurité (au cas où α_i serait corrélé). L'estimateur MCG, lui, **l'exploite,si** l'hypothèse est vraie, d'où son gain d'efficience.

Mais si l'hypothèse est fausse (comme souvent en sciences sociales), le MCG devient **biaisé** (contrairement au *within*, qui reste consistant dans tous les cas).

→ Un vrai arbitrage biais/efficience, qu'il faut trancher!

Modèles et estimateurs : la synthèse

Estimateur / vrai modèle	Pooled	Effets aléatoires	Effets fixes
Pooled OLS	Consistant	Consistant	Inconsistant
Between	Consistant	Consistant	Inconsistant
Within	Consistant	Consistant	Consistant
Différences premières	Consistant	Consistant	Consistant
MCG EA	Consistant	Consistant	Inconsistant

L'estimateur **within** est **toujours consistant**, quel que soit le vrai modèle, mais pas nécessairement le plus **efficient**.

L'estimateur des **effets aléatoires** est **inconsistant** si le vrai modèle est à effets fixes, mais **consistant et le plus efficient** si le vrai modèle est bien à effets aléatoires.

Plan

Deux façons de choisir

Le tableau précédent montre que la consistance dépend d'une hypothèse qu'on ne peut jamais vérifier directement. Deux approches, complémentaires plutôt que concurrentes :

1. Le raisonnement économique (a priori)

Avant même d'estimer quoi que ce soit : est-ce que je **soupçonne** une hétérogénéité inobservée corrélée à x_{it} ?
Un bon économètre justifie son choix de modèle par la théorie, pas seulement par un test.

2. Les tests statistiques (a posteriori)

Une fois les modèles estimés, des tests formels permettent d'**arbitrer objectivement** entre eux à partir des données : test F (pooled vs FE), test de Breusch-Pagan (pooled vs EA), test de Hausman (FE vs EA).

Le test F

On teste si les effets individuels α_i sont **conjointement nuls**, est-ce qu'aucun individu n'a de **spécificité propre**, au-delà d'une constante commune à tout le panel ?

Hypothèses

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_N = 0 \quad \text{(le pooled suffit)}$$

$$H_1 : \text{au moins un } \alpha_i \neq 0 \quad \text{(effets fixes nécessaires)}$$

On compare la somme des carrés des résidus du modèle pooled et du modèle within, si les effets fixes améliorent significativement l'ajustement, on rejette H_0.

En pratique : il existe presque toujours une hétérogénéité individuelle. Il sert surtout à **confirmer** ce qu'on soupçonnait déjà.

Le test de Breusch-Pagan (multiplicateur de Lagrange)

On teste si la variance de l'effet individuel est nulle, autrement dit, si le modèle à effets aléatoires apporte quelque chose par rapport au pooled.

Hypothèses

$$H_0 : \sigma_\alpha^2 = 0 \quad \text{(le pooled suffit)}$$

$$H_1 : \sigma_\alpha^2 > 0 \quad \text{(effets aléatoires nécessaires)}$$

Si H_0 est vraie, $\theta = 0$ (revoir la formule vue plus tôt) : le MCG se réduit exactement au pooled MCO. Le test vérifie donc si cette réduction est justifiée par les données.

Comme pour le test F : en pratique, H_0 est presque toujours rejetée dès qu'on a un vrai panel.

Le test de Hausman

On teste si l'effet individuel α_i est corrélé avec x_{it} : c'est ce qui distingue effets fixes et effets aléatoires.

Hypothèses

$$H_0 : \mathbb{E}[\alpha_i \mid x_{it}] = 0 \quad \text{(EA consistant et efficient)}$$

$$H_1 : \mathbb{E}[\alpha_i \mid x_{it}] \neq 0 \quad \text{(EA inconsistant, EF nécessaire)}$$

Sous H_0, EF et EA sont tous deux consistants, leurs coefficients devraient être **proches**. Sous H_1, seul EF reste consistant, les coefficients devraient **diverger**.

Intuition : le test compare directement $\widehat{\beta}_{EF}$ et $\widehat{\beta}_{EA}$, un grand écart entre les deux est le signe que l'hypothèse d'indépendance des effets aléatoires est probablement fausse.

Plan

Le panel utilisé pour les applications R combine plusieurs BDD (Our World in Data : OCDE, OMS, Banque Mondiale), de 2000 à 2023.

Variable	**Description**
`esperance_vie`	Espérance de vie à la naissance (années)
`depenses_sante`	Dépenses de santé par habitant ($) principal
`pib_hab`	PIB par habitant ($ constants)
`part_urbaine`	% de population vivant en zone urbaine
`pm25`	Exposition moyenne aux PM2.5 ($\mu g/m^3$)
`population`	Population totale du pays
`ratio_dependance_agee`	Ratio dépendance des personnes âgées (65+/15-64)

Deux façons d'obtenir cette base
1. Le script R, télécharge et fusionne automatiquement les données.
2. Le fichier `base_panel_sante.csv` est disponible sur Moodle.

Aperçu de la base

Ce que vous devriez obtenir avec `pdim(pdata)` :

```
Balanced Panel: n = 37, T = 24, N = 888
```

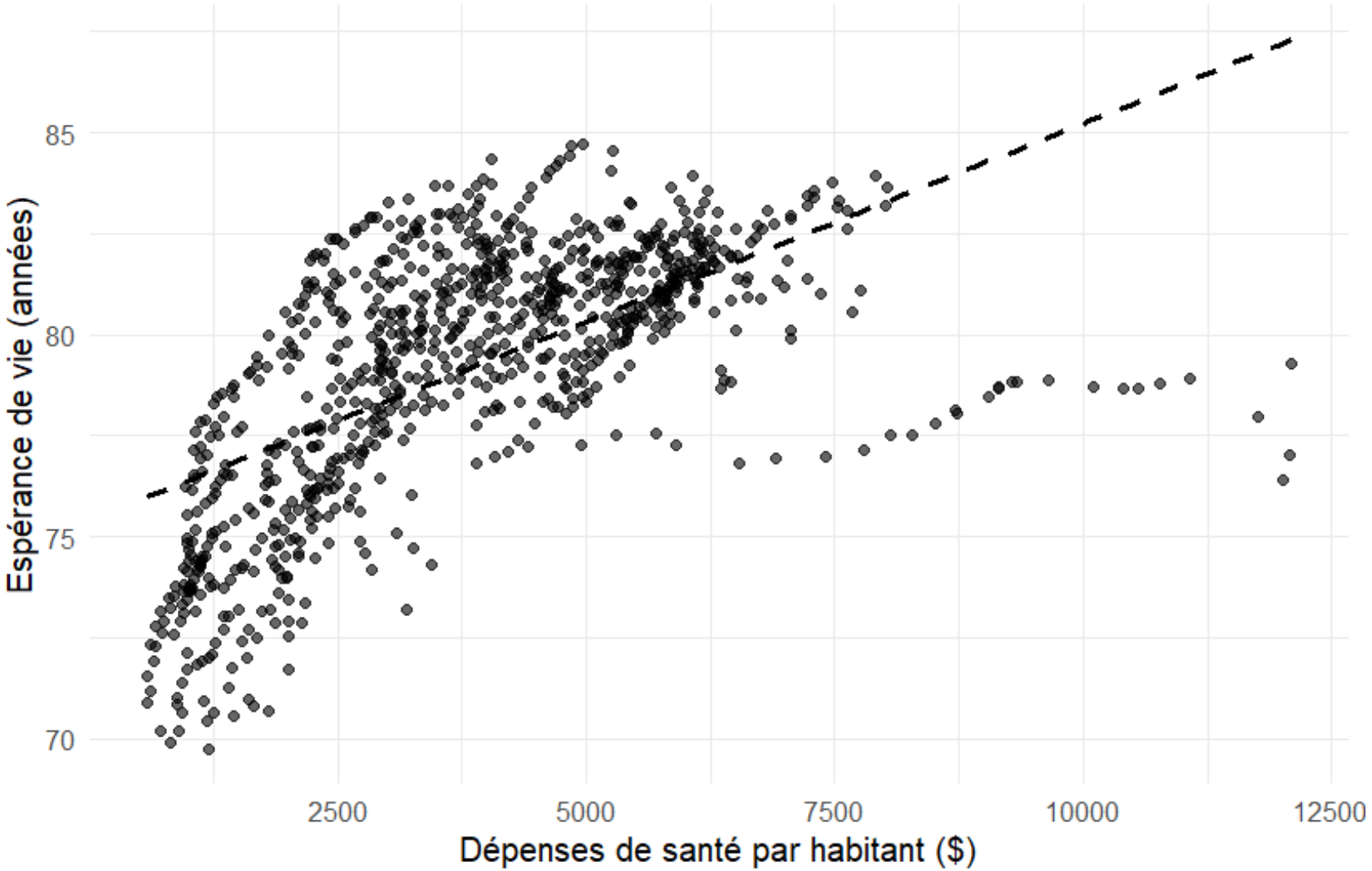
888 observations = 37 pays OCDE × 24 années (2000-2023), panel cylindré, aucune valeur manquante.

Filter

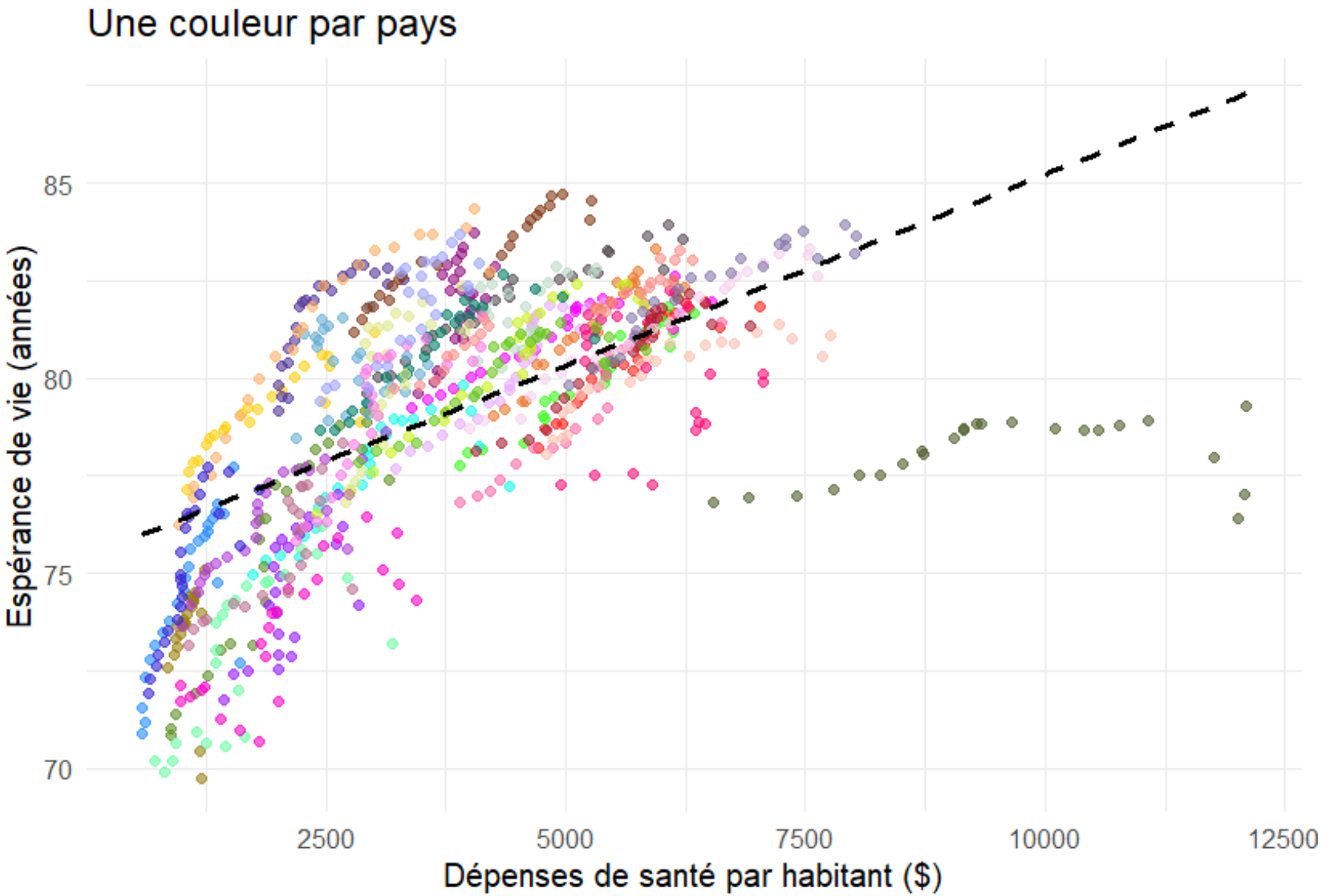
	pays	iso3c	annee	depenses_sante	esperance_vie	pib_hab	part_urbaine	pm25	population	ratio_dependance_agee
Austria-2015	Austria	AUT	2015	5925.348	81.1592	43915.227	67.55762	13.7	8644037	27.436342
Austria-2016	Austria	AUT	2016	6002.761	81.5922	44362.670	67.72181	13.0	8737920	27.537218
Austria-2017	Austria	AUT	2017	6074.392	81.6387	45056.637	67.90512	13.0	8798779	27.769827
Austria-2018	Austria	AUT	2018	6137.151	81.6859	45951.582	68.10468	14.1	8841653	28.068360
Austria-2019	Austria	AUT	2019	6258.114	81.9072	46550.562	68.31760	11.3	8880876	28.405727
Austria-2020	Austria	AUT	2020	6295.183	81.8037	43428.700	68.54102	10.2	8921402	28.831335
Austria-2021	Austria	AUT	2021	7023.391	81.8204	45368.645	68.77361	10.6	8967053	29.379332
Austria-2022	Austria	AUT	2022	6609.948	81.2959	47332.426	68.99809	10.3	9064678	30.010344
Austria-2023	Austria	AUT	2023	6457.345	81.9559	46497.910	69.22654	9.5	9130434	30.732310
Belgium-2000	Belgium	BEL	2000	3897.966	77.7684	35878.316	82.01525	14.9	10251719	25.608976
Belgium-2001	Belgium	BEL	2001	3987.216	78.0824	36148.290	82.33569	14.2	10287005	25.765669
Belgium-2002	Belgium	BEL	2002	4125.189	78.1678	36600.863	82.65096	14.5	10332975	25.883877
Belgium-2003	Belgium	BEL	2003	4571.947	78.3203	36826.270	82.96095	17.0	10376262	26.014877
Belgium-2004	Belgium	BEL	2004	4776.962	78.9605	37976.700	83.26553	14.8	10421298	26.170734

Showing 39 to 53 of 888 entries, 10 total columns

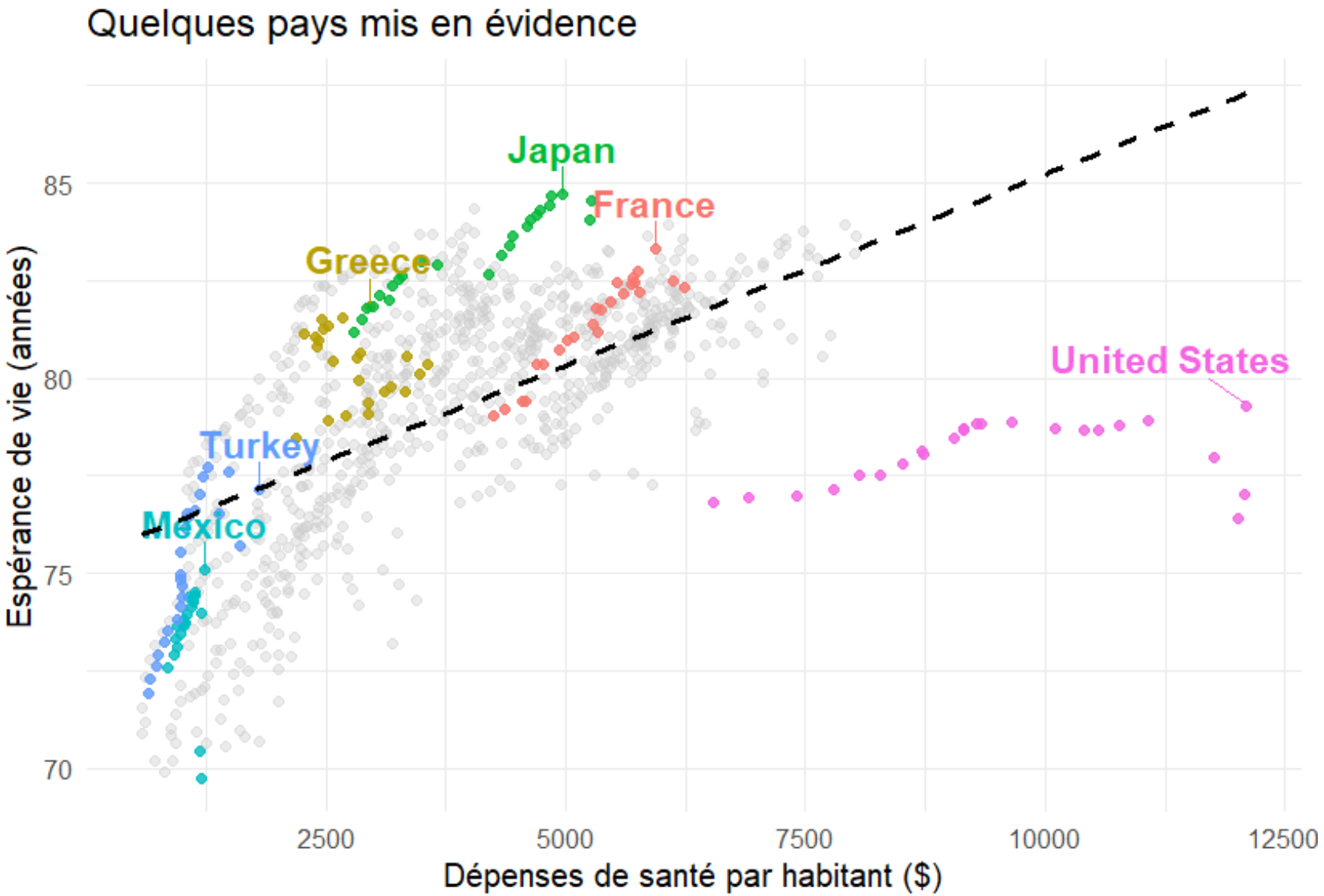
Espérance de vie et dépenses de santé



Espérance de vie et dépenses de santé



Espérance de vie et dépenses de santé



Deux packages pour estimer un modèle de panel

Il existe plusieurs packages R pour l'économétrie de panel, on en utilisera deux, complémentaires.

`plm`

Le package historique de référence en économétrie de panel sous R. Développé par Y. Croissant (Univ. Lyon 2).
Complet, bien documenté, mais plus ancien et parfois plus lent sur de gros jeux de données. Ne permet pas de faire du non linéaire, ni d'IV.

`fixest`

Package plus récent, conçu pour la rapidité (effets fixes à haute dimension) et une syntaxe compacte. Très utilisé en recherche appliquée aujourd'hui, mais uniquement en effet fixe (pas d'effets aléatoires).

Comment lire la syntaxe des commandes

plm

```
plm(y ~ x1 + x2,
data = pdata, model = "within")
```

- `y ~ x1 + x2` : variable expliquée ~ explicatives
- `data` : objet `pdata.frame` (panel déclaré)
- `model` : le modèle voulu (`"pooling"`, `"within"`, `"random"`)

fixest

```
feols(y ~ x1 + x2 | pays +
annee, data = data_complete)
```

- `y ~ x1 + x2`
- `| pays + annee` : les effets fixes à inclure (individuel, temporel, ou les deux)
- `data` : data.frame classique, pas besoin de déclarer en panel

Différence clé: `plm` choisit le modèle via l'argument `model`. `fixest` choisit le modèle via ce qu'on met **après le** `|`, pas d'effet fixe = pooled, `| pays` = effets fixes individuels, `| pays + annee` = TWFE.

Résultat : Pooled OLS

```
pooling_plm <- plm(esperance_vie ~ depenses_sante,
                   data = pdata, model = "pooling")
pooling_fixest <- feols(esperance_vie ~ depenses_sante,
                        data = data_complete)
```

Résultat : Pooled OLS

```
pooling_plm <- plm(esperance_vie ~ depenses_sante,
                   data = pdata, model = "pooling")
pooling_fixest <- feols(esperance_vie ~ depenses_sante,
                        data = data_complete)
```

```
OLS estimation, Dep. Var.: esperance_vie
Observations: 888
Standard-errors: IID
                Estimate Std. Error  t value  Pr(>|t|)
(Intercept)    75.405719   0.184464 408.7839 < 2.2e-16 ***
depenses_sante  0.000983   0.000043  22.9470 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 2.51268   Adj. R2: 0.372064
```

Résultat : Pooled OLS

```
OLS estimation, Dep. Var.: esperance_vie
Observations: 888
Standard-errors: IID
                 Estimate Std. Error  t value  Pr(>|t|)
(Intercept)    75.405719   0.184464 408.7839 < 2.2e-16 ***
depenses_sante  0.000983   0.000043  22.9470 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 2.51268   Adj. R2: 0.372064
```

+1$ de dépenses de santé → +0,00098 année d'espérance de vie (soit +0,98 an pour +1000$).
Mélange variation *between* et *within* → biaisé si α_i corrélé avec x_{it}.

Résultat : estimateur *between*

```
between_plm <- plm(esperance_vie ~ depenses_sante,
                   data = pdata, model = "between")


Oneway (individual) effect Between Model

Observations used in estimation: 37

                 Estimate Std. Error t-value  Pr(>|t|)
(Intercept)    7.5668e+01 8.8540e-01 85.462  < 2.2e-16 ***
depenses_sante 9.1409e-04 2.0838e-04  4.387   0.0001 ***

R-Squared: 0.35476
```

Résultat : estimateur *between*

```
between_plm <- plm(esperance_vie ~ depenses_sante,
                   data = pdata, model = "between")


Oneway (individual) effect Between Model

Observations used in estimation: 37

                 Estimate Std. Error t-value  Pr(>|t|)
(Intercept)    7.5668e+01 8.8540e-01 85.462  < 2.2e-16 ***
depenses_sante 9.1409e-04 2.0838e-04  4.387   0.0001 ***

R-Squared: 0.35476
```

Résultat : estimateur *between*

```
Oneway (individual) effect Between Model

Observations used in estimation: 37

                  Estimate Std. Error t-value  Pr(>|t|)
(Intercept)    7.5668e+01 8.8540e-01 85.462  < 2.2e-16 ***
depenses_sante 9.1409e-04 2.0838e-04  4.387   0.0001 ***

R-Squared: 0.35476
```

Le coefficient *between* est très proche du pooled: le pooled semble dominé par la variation between
Regardons la variance de la variable → La variance *between* (id) explique 79% de la variance totale de l'espérance de vie, contre 17% pour la tendance temporelle commune.

```
summary(pdata$esperance_vie)
total sum of squares: 8938.434
       id     time
0.7913800 0.1733856
```

Résultat : estimateur *within*

```
fe_plm <- plm(esperance_vie ~ depenses_sante,
              data = pdata, model = "within")


Oneway (individual) effect Within Model

Balanced Panel: n = 37, T = 24, N = 888

                 Estimate Std. Error t-value  Pr(>|t|)
depenses_sante 0.00144170 0.00004952  29.113 < 2.2e-16 ***

R-Squared:      0.49929
Adj. R-Squared: 0.47749
```

Résultat : estimateur *within*

```
Oneway (individual) effect Within Model

Balanced Panel: n = 37, T = 24, N = 888

                 Estimate Std. Error t-value  Pr(>|t|)
depenses_sante 0.00144170 0.00004952  29.113 < 2.2e-16 ***

R-Squared:      0.49929
Adj. R-Squared: 0.47749
```

Coefficient **plus élevé** que le pooled (0,00144 vs 0,00098) et que le *between* (0,00091) : une fois l'hétérogénéité entre pays neutralisée, l'effet des dépenses de santé sur l'espérance de vie est en réalité **plus fort**.

Résultat : Estimateur différences premières

```
d1_plm <- plm(esperance_vie ~ depenses_sante,
              data = pdata, model = "fd")
```

```
Oneway (individual) effect First-Difference Model
Balanced Panel: n = 37, T = 24, N = 888
Observations used in estimation: 851
Coefficients:
                  Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)     0.26305264  0.01960091 13.4204 < 2.2e-16 ***
depenses_sante -0.00074653  0.00010880 -6.8612 1.316e-11 ***

R-Squared:      0.052535      Adj. R-Squared: 0.051419
```

Résultat : Estimateur différences premières

```
Oneway (individual) effect First-Difference Model
Balanced Panel: n = 37, T = 24, N = 888
Observations used in estimation: 851
Coefficients:
                  Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)     0.26305264  0.01960091 13.4204 < 2.2e-16 ***
depenses_sante -0.00074653  0.00010880 -6.8612 1.316e-11 ***

R-Squared:      0.052535      Adj. R-Squared: 0.051419
```

Le coefficient devient **négatif** ! À l'opposé du *within* (+0,00144) :

- **Amplification de l'erreur de mesure** : les variations **annuelles** des 2 variables sont plus bruitées que leurs écarts à la moyenne sur 24 ans
- **Choc ponctuel** : 2020 (Covid) fait chuter l'espérance de vie **la même année** où les dépenses de santé bondissent. un choc de court terme accentué en 1D.

Résultat : Estimateur différences premières

```
Oneway (individual) effect First-Difference Model
Balanced Panel: n = 37, T = 24, N = 888
Observations used in estimation: 851
Coefficients:
                  Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)     0.26305264  0.01960091 13.4204 < 2.2e-16 ***
depenses_sante -0.00074653  0.00010880 -6.8612 1.316e-11 ***

R-Squared:      0.052535      Adj. R-Squared: 0.051419
```

$R^2 = 0{,}05$ seulement : très peu de variance expliquée → cohérent avec un estimateur qui capte surtout du bruit de court terme ici, pas une vraie relation structurelle.

Résultat : Within bidirectionnels (TWFE)

```
twfe_plm <- plm(esperance_vie ~ depenses_sante,
                data = pdata, model = "within",
                effect = "twoways")
twfe_fixest <- feols(esperance_vie ~ depenses_sante |
pays + annee, data = data_complete)


Twoways effects Within Model
Balanced Panel: n = 37, T = 24, N = 888

                  Estimate  Std. Error t-value Pr(>|t|)
depenses_sante -3.8402e-06  5.5237e-05 -0.0695   0.9446

R-Squared:      5.8443e-06        Adj. R-Squared: -0.072545
```

Résultat : Within bidirectionnels (TWFE)

```
Twoways effects Within Model
Balanced Panel: n = 37, T = 24, N = 888

                  Estimate  Std. Error t-value Pr(>|t|)
depenses_sante -3.8402e-06  5.5237e-05 -0.0695   0.9446

R-Squared:      5.8443e-06        Adj. R-Squared: -0.072545
```

L'effet **disparaît complètement** : coefficient quasi nul, non significatif ($p = 0{,}94$), $R^2 \approx 0$. En retirant à la fois l'hétérogénéité entre pays **et** la tendance temporelle commune, il ne reste presque plus de variation exploitable.

Résultat : Within bidirectionnels (TWFE)

```
Twoways effects Within Model
Balanced Panel: n = 37, T = 24, N = 888

                  Estimate  Std. Error t-value Pr(>|t|)
depenses_sante -3.8402e-06  5.5237e-05 -0.0695   0.9446

R-Squared:      5.8443e-06        Adj. R-Squared: -0.072545
```

L'effet **disparaît complètement** : coefficient quasi nul, non significatif ($p = 0{,}94$), $R^2 \approx 0$. En retirant à la fois l'hétérogénéité entre pays **et** la tendance temporelle commune, il ne reste presque plus de variation exploitable.

Rappel : le *between* expliquait 79% de la variance, la tendance temporelle 17%. Il ne restait qu'environ 4% de variation **idiosyncratique**. Le TWFE cherche un effet dans cette toute petite portion, et n'en trouve quasiment pas.

Résultat : Estimateur MCG effets aléatoires

```
random_plm <- plm(esperance_vie ~ depenses_sante,
                  data = pdata, model = "random")


Random Effect Model (Swamy-Arora's transformation)
Effects:
                 var std.dev share
idiosyncratic 1.098   1.048 0.169
individual    5.388   2.321 0.831
theta: 0.9082

                  Estimate Std. Error z-value  Pr(>|z|)
depenses_sante 1.4135e-03 4.8316e-05  29.255 < 2.2e-16 ***

R-Squared: 0.49135        Adj. R-Squared: 0.49078
```

Résultat : Estimateur MCG effets aléatoires

```
Random Effect Model (Swamy-Arora's transformation)
Effects:
                 var std.dev share
idiosyncratic 1.098   1.048 0.169
individual    5.388   2.321 0.831
theta: 0.9082

                  Estimate Std. Error z-value  Pr(>|z|)
depenses_sante 1.4135e-03 4.8316e-05  29.255 < 2.2e-16 ***

R-Squared: 0.49135        Adj. R-Squared: 0.49078
```

$\theta = 0{,}908$, très proche de 1 $\rightarrow$ le MCG se rapproche fortement du *within* (0,00141 vs 0,00144).
Cohérent avec la part `individual` très élevée (83,1% de la variance de l'erreur) : l'hétérogénéité entre pays domine largement, donc le MCG "fait presque comme si" c'était un modèle à effets fixes.

Ce qu'on retient de cette comparaison

Le coefficient varie de quasi nul et non significatif à positif et significatif, et change même de **signe** selon l'estimateur choisi. Ce n'est pas une anomalie, c'est la preuve concrète que **le choix du modèle change la conclusion**.

- **Pooled** ≈ **Between** : confirme que le pooled est dominé par la variance *between* ($\approx 79\%$ de la variance totale)
- **Within** ≈ **Effets aléatoires** : cohérent avec $\theta \approx 0{,}91$
- **TWFE** : l'effet s'effondre, peu de variation *idiosyncratique* restante ($\approx 4\%$) une fois retirés individu **et** temps
- **Différences premières** : change de signe, probablement du bruit de court terme (chocs) plutôt qu'une vraie relation

Ces premiers résultats semblent indiquer un effet positif entre dépenses de santé et espérance de vie. Reste à **trancher** entre ces modèles : les tests de spécification vont nous y aider

Application des tests sur nos données

```
pFtest(fe_plm, pooling_plm)
                    F test for individual effects
        F = 118.16, df1 = 36, df2 = 850, p-value < 2.2e-16

plmtest(pooling_plm, type = "bp")
         Lagrange Multiplier Test - (Breusch-Pagan)
         chisq = 6677.7, df = 1, p-value < 2.2e-16

phtest(fe_plm, random_plm)
                             Hausman Test
           chisq = 6.7524, df = 1, p-value = 0.009362
```

Les trois p-values sont $< 0{,}05$: on rejette H_0 à chaque fois.

- **Test F** et **Breusch-Pagan** : rejet **massif** de H_0 → l'hétérogénéité individuelle est indiscutable, un modèle de panel est nécessaire
- **Hausman** : rejet **plus modeste** → EF et EA divergent significativement

Conclusion du chapitre

- **Raisonnement économique + tests statistiques** convergent vers les **effets fixes**
- La **plupart des modèles** (pooled, between, within, EA) vont dans le même sens : un effet **positif** des dépenses de santé sur l'espérance de vie
- Seul le **TWFE** s'écarte : car il **manque de puissance** (trop peu de variation *idiosyncratique* restante)

Notre spécification la plus robuste : les effets fixes. +1000$ de dépenses de santé par habitant → +1,4 année d'espérance de vie.

Cette conclusion reste **provisoire** : il faut encore travailler la **validité des erreurs standards** (séance 3) et la **stratégie d'identification causale** (DiD, IV : séances 4 et 5) avant de parler d'un vrai effet causal.