Économétrie des données de panel

M2 Économie de la Santé et Économie Appliquée

Université Paris Cité

Carla Morvan

2026-2027

Part I

Données de panel

Données de panel

Plan

Objectifs du chapitre

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

- Distinguer une donnée de panel d'une coupe transversale et d'une série temporelle
- Expliquer l'intérêt majeur du panel : le contrôle de l'hétérogénéité inobservée
- Différencier la variance *within* et la variance *between*
- Déclarer et manipuler un panel sous R (formats *wide*/*long*, statistiques descriptives)

Qu'est-ce qu'une donnée de panel ?

Définition

Une donnée de panel (ou donnée longitudinale) suit les **mêmes unités** sur **plusieurs périodes** de temps.

Elle combine donc deux dimensions :

- une dimension **individuelle** i
  - $i = 1, \ldots, N$: indice **individu** (pays, patient, hôpital...)
- une dimension **temporelle** t
  - $t = 1, \ldots, T$: indice **temps** (année, mois...)

→ Nos variables vont donc être sous la forme

- y_{it} : variable d'intérêt pour l'individu i à la période t
- x_{it} : variable(s) explicative(s) pour l'individu i à la période t

Trois types de données

Type	**Dimension**	**Exemple**
Coupe transversale	i	Concentration de pollution par ville en 2020
Série temporelle	t	Nombre de lits d'hospitalisation en France, 1990-2025
Panel	i et t	Dépenses de santé de chaque pays d'Europe, 2000-2020

Le panel n'est pas juste "plus de données" : c'est une structure qui permet de suivre l'évolution **de chaque unité**.

Obs.	Individu	Temps	y	x_1	...	x_k
1	i_1	t_1	$y_{1,1}$	$x_{1,1,1}$	$\cdots$	$x_{k,1,1}$
2	i_1	t_2	$y_{1,2}$	$x_{1,1,2}$	$\cdots$	$x_{k,1,2}$
$\vdots$	i_1	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	i_1	t_T	$y_{1,T}$	$x_{1,1,T}$	$\cdots$	$x_{k,1,T}$
	i_2	t_1	$y_{2,1}$	$x_{1,2,1}$	$\cdots$	$x_{k,2,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i_3	t_1	$y_{3,1}$	$x_{1,3,1}$	$\cdots$	$x_{k,3,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N \times T$	i_N	t_T	$y_{N,T}$	$x_{1,N,T}$	$\cdots$	$x_{k,N,T}$

Chaque ligne est une observation : un individu i, à une période t, avec ses caractéristiques $y_{i,t}$ et $x_{1,i,t}, \ldots, x_{k,i,t}$.
C'est le format *long*, le plus courant pour les données de panel au total $N \times T$ lignes.

Obs.	Individu	Temps	y	x_1	...	x_k
1	i_1	t_1	$y_{1,1}$	$x_{1,1,1}$	$\cdots$	$x_{k,1,1}$
2	i_1	t_2	$y_{1,2}$	$x_{1,1,2}$	$\cdots$	$x_{k,1,2}$
$\vdots$	i_1	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	i_1	t_T	$y_{1,T}$	$x_{1,1,T}$	$\cdots$	$x_{k,1,T}$
	i_2	t_1	$y_{2,1}$	$x_{1,2,1}$	$\cdots$	$x_{k,2,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i_3	t_1	$y_{3,1}$	$x_{1,3,1}$	$\cdots$	$x_{k,3,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N \times T$	i_N	t_T	$y_{N,T}$	$x_{1,N,T}$	$\cdots$	$x_{k,N,T}$

Série temporelle : on empile les observations d'un seul individu (i_1) sur les T périodes.

Obs.	Individu	Temps	y	x_1	...	x_k
1	i_1	t_1	$y_{1,1}$	$x_{1,1,1}$	$\cdots$	$x_{k,1,1}$
2	i_1	t_2	$y_{1,2}$	$x_{1,1,2}$	$\cdots$	$x_{k,1,2}$
$\vdots$	i_1	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	i_1	t_T	$y_{1,T}$	$x_{1,1,T}$	$\cdots$	$x_{k,1,T}$
	i_2	t_1	$y_{2,1}$	$x_{1,2,1}$	$\cdots$	$x_{k,2,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i_3	t_1	$y_{3,1}$	$x_{1,3,1}$	$\cdots$	$x_{k,3,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N \times T$	i_N	t_T	$y_{N,T}$	$x_{1,N,T}$	$\cdots$	$x_{k,N,T}$

Coupe transversale : on garde une seule période (t_1) pour les N individus.

Obs.	Individu	Temps	y	x_1	...	x_k
1	i_1	t_1	$y_{1,1}$	$x_{1,1,1}$	$\cdots$	$x_{k,1,1}$
2	i_1	t_2	$y_{1,2}$	$x_{1,1,2}$	$\cdots$	$x_{k,1,2}$
$\vdots$	i_1	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	i_1	t_T	$y_{1,T}$	$x_{1,1,T}$	$\cdots$	$x_{k,1,T}$
	i_2	t_1	$y_{2,1}$	$x_{1,2,1}$	$\cdots$	$x_{k,2,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i_3	t_1	$y_{3,1}$	$x_{1,3,1}$	$\cdots$	$x_{k,3,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N \times T$	i_N	t_T	$y_{N,T}$	$x_{1,N,T}$	$\cdots$	$x_{k,N,T}$

Panel : on empile les N individus, sur les T périodes : $N \times T$ lignes au total.

Pourquoi s'y intéresser en économie appliquée ?

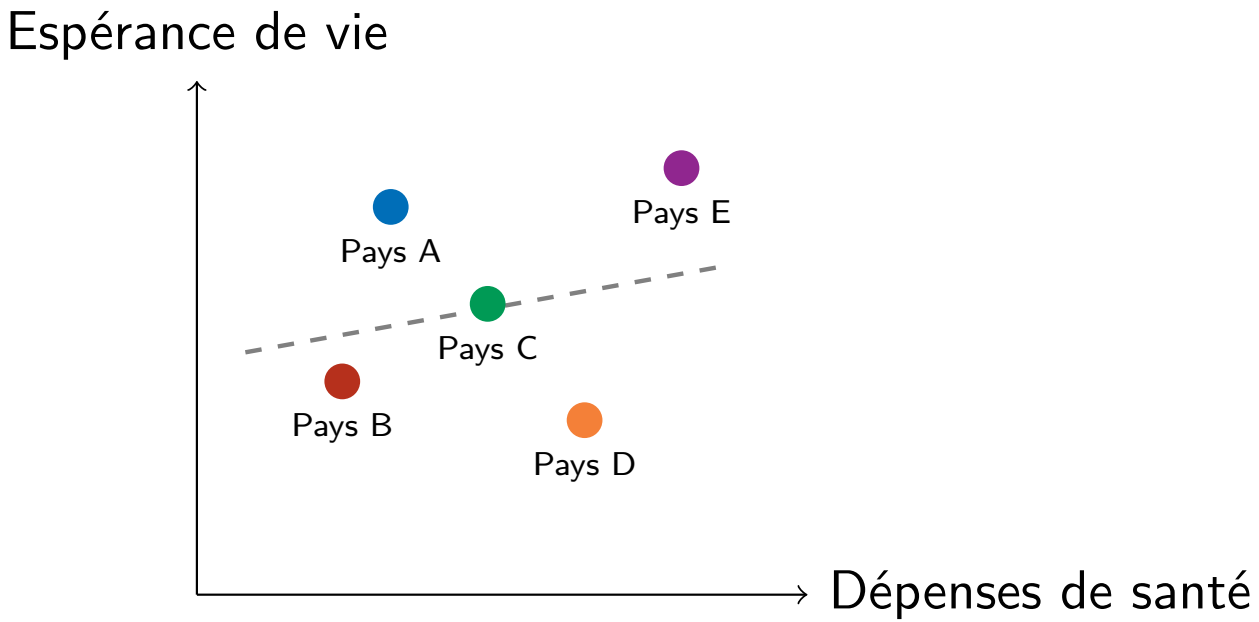
- La très grande majorité des **évaluations de politiques publiques** (santé, environnement, etc...) s'appuient sur des données de panel
- C'est le cadre naturel pour répondre à une question centrale en économie appliquée : **corrélation ou causalité ?**
- Les méthodes de ce cours (effets fixes, DiD, IV en panel) sont **omniprésentes** dans la littérature empirique récente → car le panel est un outil très puissant !

Plan

Le problème sans panel

Exemple

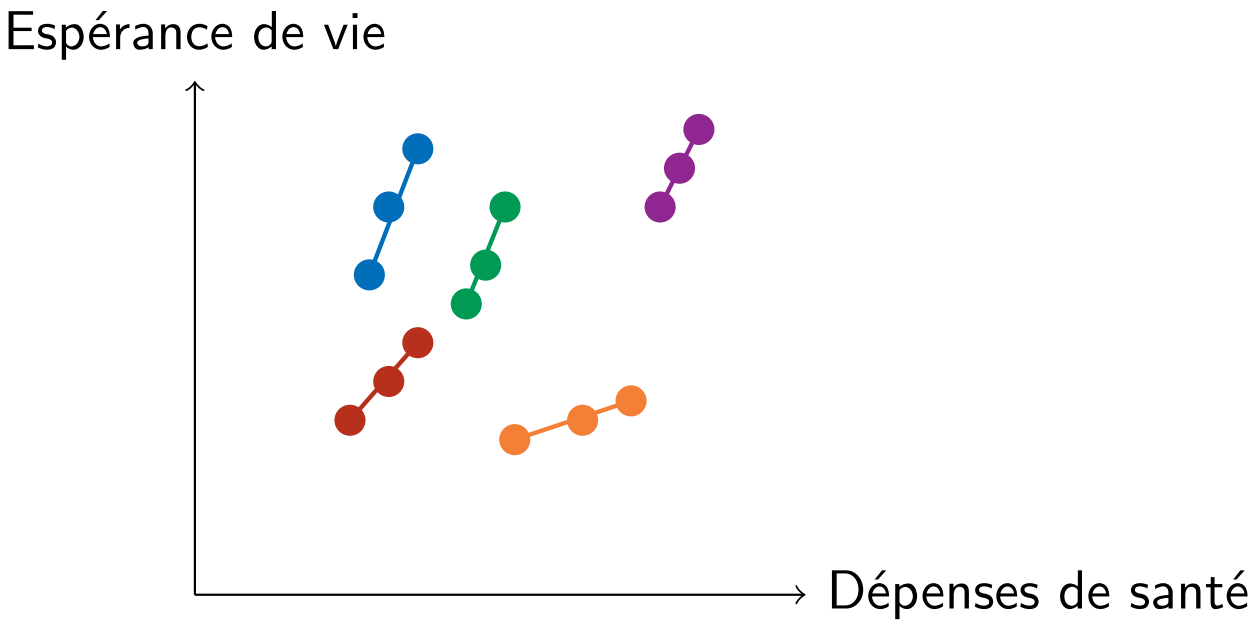
On veut étudier l'effet des dépenses de santé sur l'espérance de vie.



→ **Problème :** Si un pays dépense plus **et** a une espérance de vie plus élevée, est-ce parce que les dépenses de santé sont efficaces... ou parce que ce pays est simplement plus riche, plus développé, avec de meilleures infrastructures ?

Ce que permet le panel

Avec des données de panel, on observe **chaque pays plusieurs fois** dans le temps.



Tout ce qui caractérise un pays et **ne change pas (ou peu) dans le temps** — culture, institutions, géographie... — peut être **neutralisé**, puisqu'on compare le pays **à lui-même** à différentes périodes.

⇒ **contrôle de l'hétérogénéité inobservée *invariante dans le temps***

Within vs Between

Pour une variable y_{it} (l'individu i à la période t), on peut décomposer sa variation totale en deux sources.

Moyenne individuelle

$$\overline{y}_i = \frac{1}{T} \sum_{t=1}^{T} y_{it}$$

la moyenne de y pour l'individu i, sur toute la période T.

Variance *between*

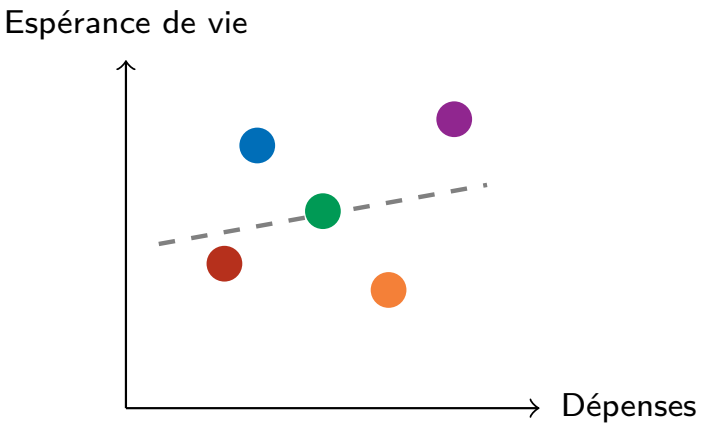
Variation **entre individus** : comment $\overline{y}_i$ diffère d'un individu à l'autre

Variance *within*

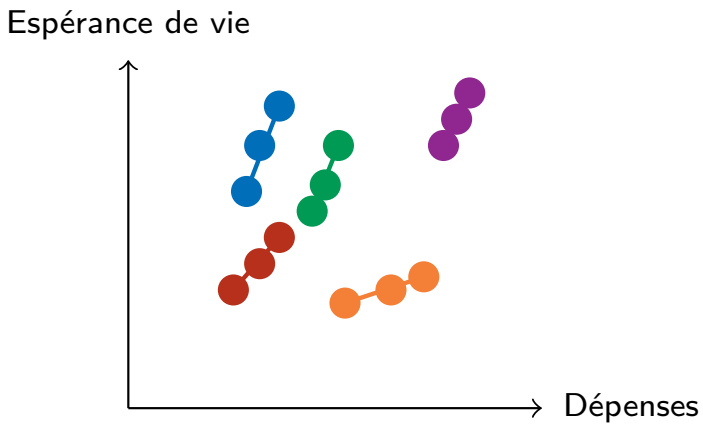
Variation **dans le temps, pour un même individu** : l'écart par rapport à sa propre moyenne $y_{it} - \overline{y}_i$

Within vs Between

- Le graphique **"sans panel"** (coupe transversale) ne capture que la variation **between** : comparer des pays différents
- Le graphique **"avec panel"** exploite en plus la variation **within** : comparer un pays à lui-même



Between



Within

C'est cette variation *within* qui permet de neutraliser l'hétérogénéité inobservée invariante dans le temps.

Panel cylindré vs non cylindré

Panel cylindré (*balanced*)

Chaque individu i est observé à **toutes** les périodes $t = 1, \ldots, T$. Le panel a exactement $N \times T$ observations.

Panel non cylindré (*unbalanced*)

Certains individus ont des observations **manquantes** à certaines périodes (entrée tardive, sortie prématurée, données non disponibles). Le nombre d'observations est **inférieur** à $N \times T$.

Exemple santé : un hôpital qui ferme en cours de période, ou un patient perdu de vue dans une cohorte.

L'attrition : un cas particulier à surveiller

Définition

L'attrition désigne la **sortie progressive** d'individus du panel au cours du temps (abandon, décès, faillite, déménagement...).

Pourquoi c'est un problème ?

Si les individus qui quittent le panel ne le font **pas au hasard** (ex. : les patients qui vont le moins bien arrêtent le suivi)
→ les estimations peuvent être **biaisées** : ce n'est plus un problème de taille d'échantillon, mais un **problème de sélection** !

→ Un panel non cylindré n'est pas automatiquement un problème, mais qu'il faut se demander **pourquoi** il l'est.

Plan

Importer un panel sous R

On utilise un panel de dépenses de santé, mis à disposition sur Moodle (`depenses_sante_ocde.csv`), construit à partir de **Our World in Data** (OCDE/OMS, via Banque Mondiale).

Our World in Data (OCDE/OMS, Banque Mondiale)

ourworldindata.org est une excellente source pour trouver des panels prêts à l'emploi (santé, environnement, économie...).

Importer un panel sous R

On utilise un panel de dépenses de santé, mis à disposition sur Moodle (`depenses_sante_ocde.csv`), construit à partir de **Our World in Data** (OCDE/OMS, via Banque Mondiale).

Our World in Data (OCDE/OMS, Banque Mondiale)

ourworldindata.org est une excellente source pour trouver des panels prêts à l'emploi (santé, environnement, économie...).

```
library(tidyverse)
```

Rappel : `tidyverse` est une collection de packages R (`dplyr`, `ggplot2`, `tidyr`...) pour manipuler et visualiser des données.

Importer un panel sous R

On utilise un panel de dépenses de santé, mis à disposition sur Moodle (`depenses_sante_ocde.csv`), construit à partir de **Our World in Data** (OCDE/OMS, via Banque Mondiale).

Our World in Data (OCDE/OMS, Banque Mondiale)

ourworldindata.org est une excellente source pour trouver des panels prêts à l'emploi (santé, environnement, économie...).

```
library(tidyverse)
```

Rappel : `tidyverse` est une collection de packages R (`dplyr`, `ggplot2`, `tidyr`...) pour manipuler et visualiser des données.

```
data <- read_csv("data_depenses_sante_ocde.csv")
head(data)
str(data)
```

Un panel peut être stocké sous deux formes :

Wide

Une ligne par individu, une colonne par période. → lisible pour un humain, pratique pour **Excel**

Long

Une ligne par observation (i, t). Format attendu par les packages **économétriques**.

Un panel peut être stocké sous deux formes :

Wide

Une ligne par individu, une colonne par période. → lisible pour un humain, pratique pour **Excel**

Long

Une ligne par observation (i, t). Format attendu par les packages **économétriques**.

```
# wide -> long
data_long <- data %>%
  pivot_longer(cols = starts_with("depenses_sante_"),
               names_to = "annee",
               names_prefix = "depenses_sante_",
               values_to = "depenses")
# long -> wide
data_wide <- data_long %>%
  pivot_wider(names_from = annee,
              values_from = depenses,
              names_prefix = "depenses_sante_")
```

Déclarer un panel sous R

Il faut indiquer à R quelles colonnes identifient l'individu et le temps.

```
library(plm)
```

> `plm` est un package R qui permet de manipuler et visualiser des données de panel.

Déclarer un panel sous R

Il faut indiquer à R quelles colonnes identifient l'individu et le temps.

```
library(plm)
```

`plm` est un package R qui permet de manipuler et visualiser des données de panel.

```
pdata <- pdata.frame(data_long,
                     index = c("pays", "annee"))

# Vérifier la structure du panel
pdim(pdata)
```

Ce que renvoie `pdim()`

```
Balanced Panel: n = 37, T = 21, N = 777
```

Le nombre d'individus n, le nombre de périodes T, et le nombre d'observation N.

Statistiques descriptives within/between

```
summary(pdata$depenses)
```

```
total sum of squares: 2821023000
        id       time
0.89722407 0.06929389

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    592    2186    3567    3687    5016   12077
```

`id` : part de la variance totale expliquée par les **différences entre pays** (variation *between*) $\approx 90\%$.
`time` : part expliquée par une **tendance temporelle commune** $\approx 7\%$.
Le reste ($\approx 3\%$) est la variation résiduelle, propre à chaque pays dans le temps.

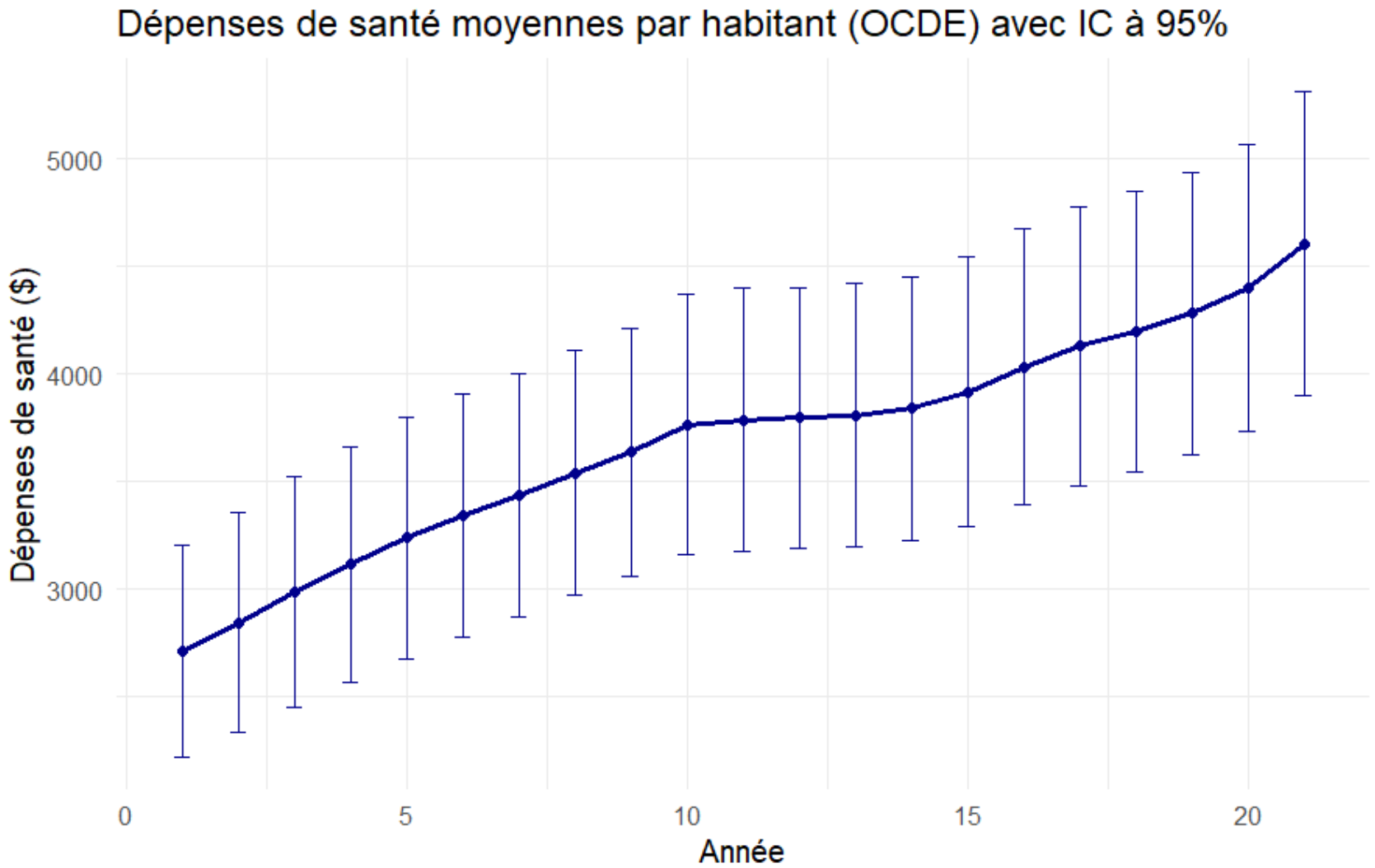
Visualiser le panel : la moyenne par année

L'évolution de la **moyenne du panel**, tous individus confondus.

```
pdata %>%  group_by(annee) %>%
  summarise(moy = mean(depenses, na.rm = TRUE),
            ecart_type = sd(depenses, na.rm = TRUE), n = n(),
            se = ecart_type / sqrt(n),
            ic_bas = moy - 1.96 * se, ic_haut = moy + 1.96 * se) %>%
  ggplot(aes(x = as.numeric(annee), y = moy)) +
  geom_errorbar(aes(ymin = ic_bas, ymax = ic_haut),
  color = "darkblue", width = 0.3) +
  geom_line(color = "darkblue", linewidth = 1) +
  geom_point(color = "darkblue", size = 1.5) +
  labs(x = "Année", y = "Dépenses de santé ($)",
  title = "Dépenses de santé moyennes par habitant (OCDE) avec IC à 95%
  theme_minimal()
```

Attention à l'interprétation: Cette moyenne écrase **toute la variation *between*** : elle ne dit rien des écarts entre pays, seulement la tendance **agrégée** dans le temps.

Visualiser le panel : la moyenne par année



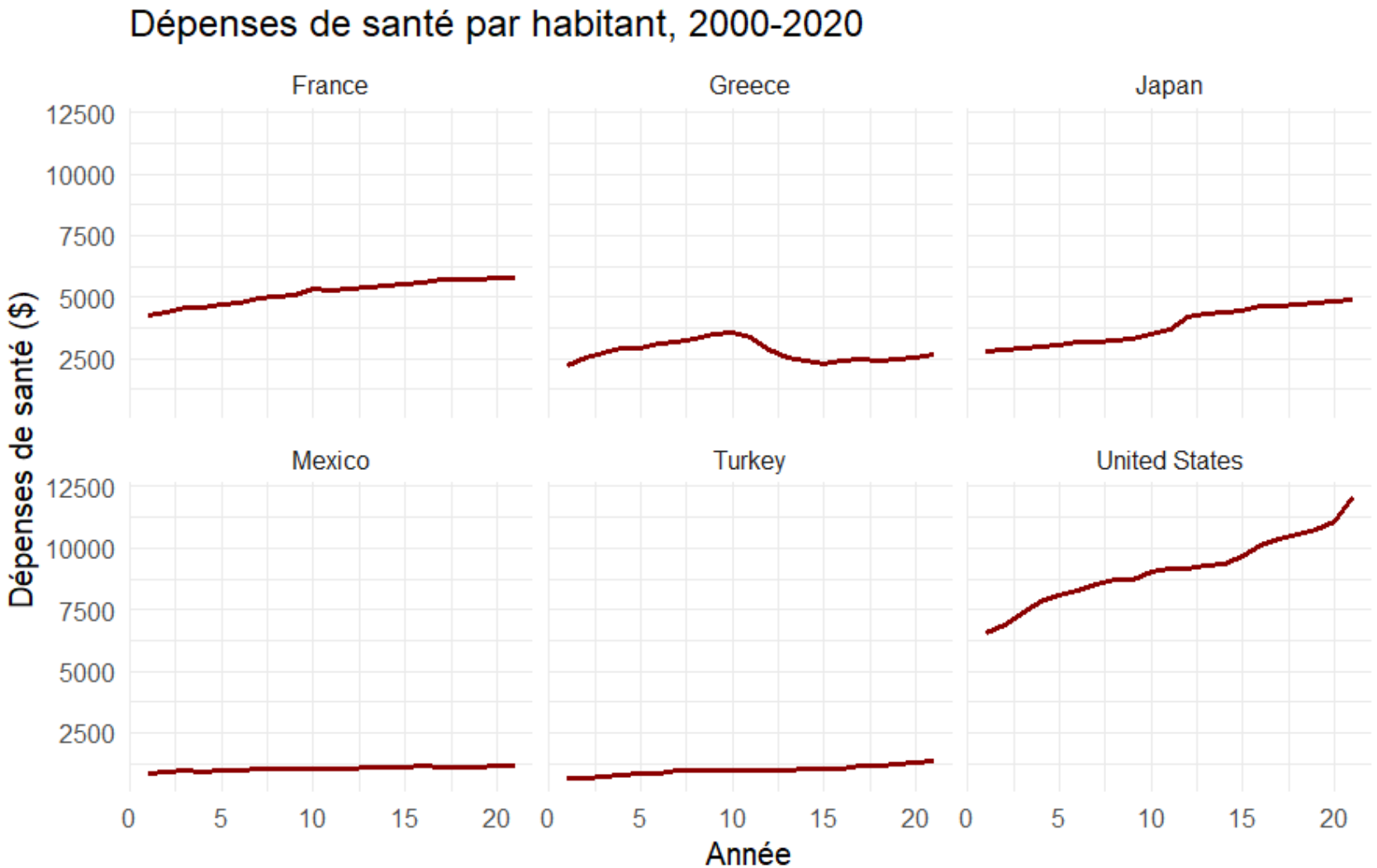
Visualiser le panel : `facet_wrap`

L'évolution de chaque individu séparément.

```
pdata %>%
  filter(pays %in% c("France", "United States","Japan",
                     "Mexico", "Turkey", "Greece")) %>%
  ggplot(aes(x = as.numeric(annee), y = depenses)) +
  geom_line(color="darkred", linewidth = 1) +
  facet_wrap(~ pays) +
  labs(x = "Année", y = "Dépenses de santé ($)",
       title = "Dépenses de santé par habitant, 2000-2020") +
  theme_minimal()
```

→ On visualise directement la variation ***within***: chaque panneau montre la trajectoire propre d'un pays, indépendamment des niveaux des autres.

Visualiser le panel : `facet_wrap`



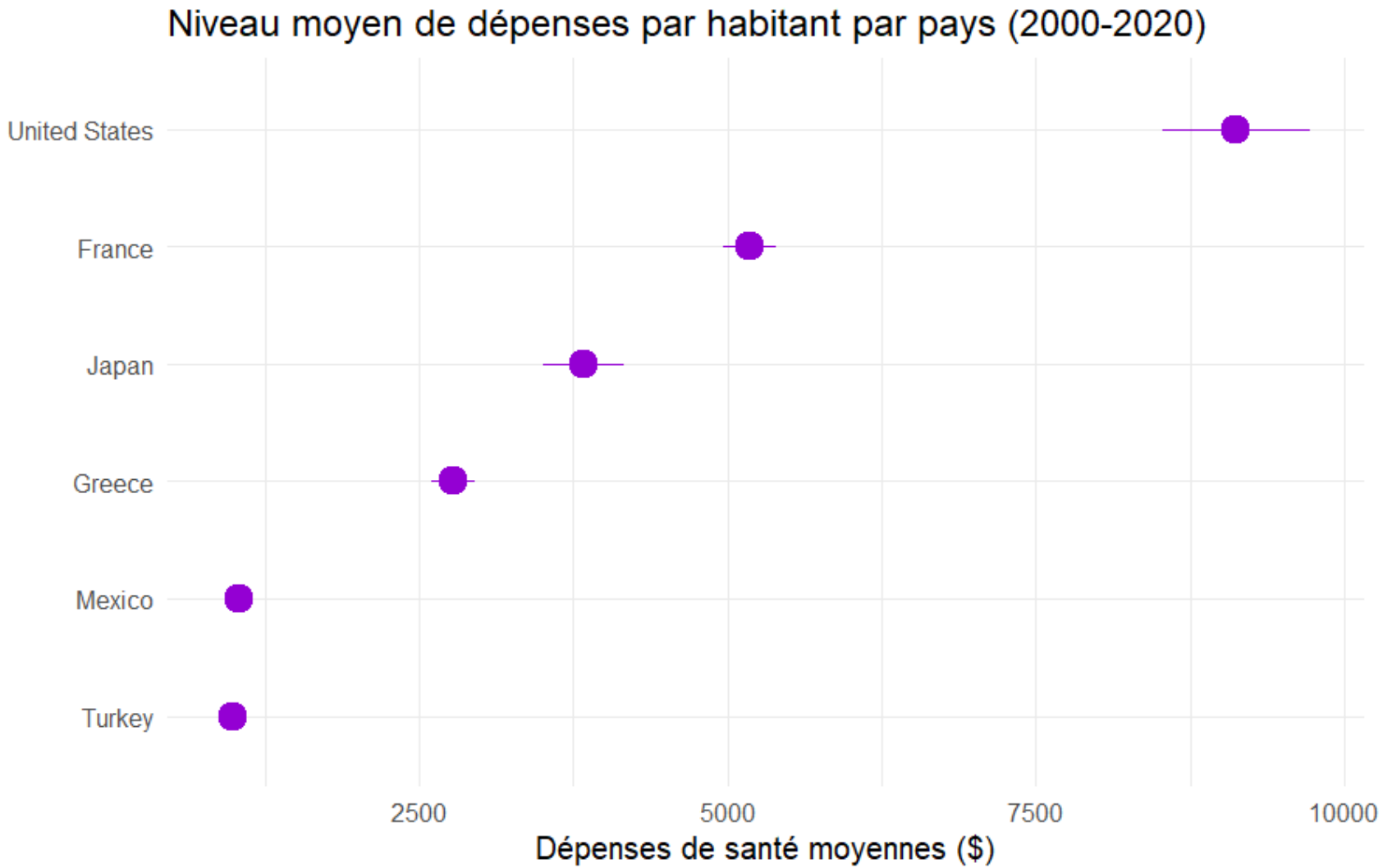
Visualiser le panel : Niveau moyen par pays

On regarde la **moyenne dans le temps, pour chaque pays**, la variation *between*.

```
pdata %>%  group_by(pays) %>%
  filter(pays %in% c("France", "United States",  "Japan",
  "Mexico", "Turkey", "Greece")) %>%
  summarise(moy = mean(depenses, na.rm = TRUE),
            se = sd(depenses, na.rm = TRUE) / sqrt(n()),
            ic_bas = moy - 1.96 * se, ic_haut = moy + 1.96 * se) %>%
  ggplot(aes(x = reorder(pays, moy), y = moy)) +
  geom_pointrange(aes(ymin = ic_bas, ymax = ic_haut),
                  color = "darkviolet", size = 1) + coord_flip() +
  labs(x = NULL, y = "Dépenses de santé moyennes ($)",
       title = "Niveau moyen de dépenses par habitant par pays (2000-20
  theme_minimal()
```

L'écart de niveau entre pays (ex. Turquie $\approx$ 1000\$ vs États-Unis $\approx$ 9100\$) domine très largement la marge d'erreur de chaque estimation. Cohérent avec la part *between* $\approx$ 90% vue dans `summary()`.

Visualiser le panel : Niveau moyen par pays



Récapitulatif : notre panel

- **Panel cylindré** : $n \times T = 37 \times 21 = 777 = N \rightarrow$ aucune observation manquante
- **Pas d'attrition** : tous les pays sont suivis sur l'intégralité de la période 2000-2020
- **Variance *between* $\approx$ 90%** : les pays diffèrent énormément entre eux (niveau de richesse, infrastructures...)
- **Variance *time* $\approx$ 7%** : peu de tendance commune à tous les pays

$\rightarrow$ Un panel propre et cylindré, mais dominé par la variation *between*
$\rightarrow$ un terrain idéal pour illustrer l'intérêt des **effets fixes** : neutraliser cette énorme hétérogénéité entre pays pour isoler la variation *within*