Econométrie des panels

Olivier Godechot

Panel: exemple

- Waldfogel Jane,
 1997, "The effect of Children on woman's wage",
 ASR
- Qu'est-ce qui favorise la pénalité salariale liée aux enfants ?
 - Hétérogénéité inobservée : Non
 - Temps de travail incomplet : Oui

Table 4. Coefficients from Pooled OLS Models and Fixed-Effects Models Regressing Ln Hourly Wage on Selected Family Status Variables: Women from the NLS-YW, 1968–1988

	Poo	led OLS Mod	lels	Fixe	d-Effects Mo	dels
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Actual work experience	.025* (.003)	.025* (.002)		.027* (.002)	.025* (.002)	_
Actual work experience squared	000 (.000)	000 (.000)	_	000 (.000)	000* (.000)	_
Part-time work experience		_	002 (.004)	_	_	.024* (.003)
Part-time work experience squared		_	.002* (.000)	_	_	000* (.000)
Full-time work experience		_	.029* (.003)	_	_	.025* (.002)
Full-time work experience squared		_	000 (.000)	_		000 (.000)
Age	.110* (.007)	.093* (.007)	.095* (.006)	.112* (.003)	.104* (.003)	.104* (.003)
Age squared	002* (.000)	002* (.000)	002* (.000)	002* (.000)	002* (.000)	002* (.000)
Education	.070* (.002)	.069* (.002)	.069* (.002)	.059* (.002)	.051* (.002)	.051* (.002)
Married	.056* (.011)	.047* (.010)	.045* (.010)	.042* (.007)	.034* (.006)	.034* (.006)
Divorced	.088* (.015)	.065* (.015)	.064* (.014)	.062* (.009)	.048* (.009)	.047* (.009)
One child	055* (.012)	043* (.011)	041* (.011)	056* (.006)	039* (.006)	038* (.006)
Two or more children	133* (.013)	107* (.013)	096* (.013)	147* (.008)	117* (.008)	116* (.008)
Children not in the home	034* (.005)	033* (.005)	027* (.005)	033* (.003)	030* (.003)	029* (.003)
Part-time work currently	-	145* (.010)	123* (.010)	_	111* (.005)	112* (.005)
Black	029* (.010)	042* (.010)	045* (.010)	-	_	_
Hispanic	.040 (.023)	.040 (.022)	.040 (.022)	_	_	_

Panel (Définition)

- On parle de données de panel lorsque l'on dispose pour le même individu de plusieurs observations à plusieurs dates.
- On oppose les données de panel (ou longitudinales) et données en coupe (cross-sectional)
- Intérêt du panel
 - Évolutions individuelles
 - Estimation des paramètres sur la base non seulement des différences interindividuelles mais aussi des différences intra-individuelles
 - Distinction des effets d'âge et de génération
 - Possibilité d'introduire des effets retards.
 - Permet d'établir de manière plus convaincante des relations de causalité

Panel (introduction - suite)

- Quelques panels
 - Enquête emploi,
 - -> 2002 données panélisées sur trois ans avec renouvellement par tiers (trois interrogations)
 - Après 2002, données panélisées sur 18 mois avec renouvellement par sixième (six interrogations)
 - Panel européen des ménages (1994-2001)
 - Questions sur les relations sociales, la santé
 - Statistiques sur les ressources et les conditions de vie (SRCV) (2004-.).
 - Panel de 9 ans renouvelé par 1/9ème avec une interrogation par an.
 - Panel des élèves du second degré
 - Échantillon démographique permanent
 - Panel des DADS

Organisation des données

- Données empilées (le plus classique)
- Pour le package plm de R (Panel)
 - [Anciens pré-requis] Mettre l'identifiant individuel en première colonne et le temps en deuxième colonne

	ident	an	S	SALRED	нн	salhor	ехр	age	DDIPL	M	ENFC90	ENF3
1	7213010300010101	1999	1	14000.000	35	92.59259	24	44	5	1	0	0
2	7213010300010101	2000	1	12000.000	35	79.36508	25	45	5	1	0	0
3	7213010300010101	2001	1	13000.000	35	85.97884	26	46	5	1	0	0
4	7213010300180101	1999	1	5000.000	39	29.67711	1.6	30	7	2	0	0
5	7213010300180101	2000	1	5667.000	35	37.48016	1	31	7	2	1	1
6	7213010300180101	2001	1	5981.167	35	39.55798	2	32	7	2	1	1
7	7213010300180102	2001	2	6000.000	35	39.68254	0.4	29	3	2	1	1
8	7213010300230101	1999	1	5500.000	26	48.96724	4	43	3	2	3	0
9	7213010300230101	2000	1	6000.000	19	73.09942	6	44	3	2	3	0
10	7213010300230102	1999	2	8500.000	26	75.67664	0.33	39	4	2	3	0
11	7213010300230102	2000	2	10000.000	35	66.13757	1.5	40	4	2	3	0
12	7213010300230102	2001	2	10000.000	26	89.03134	2	41	4	2	3	0

Organisation des données

• En colonnes (plus rares)

	ident	S	SALRED.1999	SALRED.2000	SALRED.2001	HH.1999	HH.2000	HH.2001
1	7213010300010101	1	14000.000	12000.000	13000.000	35	35	35
2	7213010300180101	1	5000.000	5667.000	5981.167	39	35	35
3	7213010300180102	2	NA	NA	6000.000	NA	NA	35
4	7213010300230101	1	5500.000	6000.000	NA	26	19	NA
5	7213010300230102	2	8500.000	10000.000	10000.000	26	35	26

- Multiplication du nombre de colonnes surtout si la dimension temps importante
- Peu pratique pour certaines manipulations
- D'autres calculs plus faciles

Les panels comme problème et comme solution

Organisation des données (fin)

- Panel cylindré (balanced)
 - − => même nombre d'observations par individus.
- Panel non cylindré (unbalanced)
 - données manquantes, disparition du panel, entrée dans le panel, etc.
- Les résultats de l'économétrie des panels d'abord établis pour les panel cylindrés, puis extension aux panel noncylindrés.

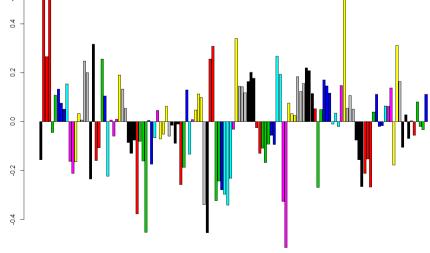
Problèmes et solutions : autocorrélation des résidus

• Les deux problèmes posés par l'économétrie des panels

$$y_{it} = \beta_0 + \beta_1 x_{1it} + ... + \beta_k x_{kit} + u_{it}$$

- Problème n°1: autocorrélation des résidus.
 - $-\operatorname{Cov}(u_{it}, u_{it+1}) = 0$?

 Si je suis sous-payé à la date t, je le serai sans doute à la date t+1



Hétérogénéité inobservée

• Le problème d'hétérogénéité ou des variables confondantes se pose en panel comme en MCO classique

$$y_{it} = \beta_0 + \beta_1 x_{1it} + ... + \beta_k x_{kit} + u_{it}$$

Si $u_{it} = c.inobs + e$ et $cov(x_{ik}, inobs) \neq 0$... alors biais!

- Le résidu u_{it} peut être réécrit de la façon suivante : $u_{it} = a_i + e_{it}$, comme la somme d'une erreur individuelle constante a_i , et e_{it} une erreur temporaire.
- *a_i* peut être vu comme déterminé par les variables inobservées constantes dans le temps

Hétérogénéité inobservée (2)

• Question : peut-on dire que dans notre vrai modèle cette erreur constante a_i par individu est indépendante des variables explicatives?

```
soit cov(a_i, x_k) = 0?
```

- Si oui, l'estimation pooling ou aléatoire est bonne
- Si non, alors l'hypothèse des MCO $cov(u_i, x_k) = 0$ n'est pas respectée
- Le modèle des MCO n'est pas consistant pour estimer les $oldsymbol{eta}$
- Mais on peut traiter ici le problème avec les modèles à effets fixes ou en différences premières

Combiner la variance intra-individuelle et la variance interindividuelle

... et régler le problème de l'auto-corrélation des résidus

Le modèle homogène

• On peut estimer l'équation suivante avec les MCO.

$$y_{it} = \beta_0 + \beta_1 x_{1it} + ... + \beta_k x_{kit} + u_{it}$$

- « Pooled ou pooling regression » ou modèle homogène
 - Soit tout simplement avec le modèle lm

- Soit avec le package plm

Exemple salaire et enfants

```
Oneway (individual) effect Pooling Model
Unbalanced Panel: n=10679, T=1-3, N=20450
              Estimate Std. Error t-value Pr(>|t|)
(Intercept) 3.2927e+00 8.4863e-02 38.8005 < 2.2e-16 ***
            2.2499e-02 1.1765e-03 19.1236 < 2.2e-16 ***
exp
           -2.4025e-04 4.4576e-05 -5.3898 7.132e-08 ***
exp2
           4.1935e-02 4.7122e-03 8.8993 < 2.2e-16 ***
age
age2
           -4.5960e-04 6.1953e-05 -7.4185 1.231e-13 ***
           -2.4164e-01 1.0152e-02 -23.8018 < 2.2e-16 ***
DDIPL3
           -4.4744e-01 1.0222e-02 -43.7716 < 2.2e-16 ***
DDIPL4
           -6.1445e-01 9.4473e-03 -65.0394 < 2.2e-16 ***
DDIPL5
           -6.0920e-01 1.2280e-02 -49.6070 < 2.2e-16 ***
DDIPL6
           -7.5569e-01 1.0135e-02 -74.5631 < 2.2e-16 ***
DDIPL7
           -1.8977e-02 8.0101e-03 -2.3691 0.017842 *
M2
           -3.6286e-02 2.5193e-02 -1.4403 0.149789
M3
           3.3106e-02 1.0946e-02 3.0246 0.002493 **
М4
           -6.9214e-03 8.2204e-03 -0.8420 0.399814
COHAB2
immiq
           4.1933e-03 9.4416e-03 0.4441 0.656952
           -6.0055e-03 2.8873e-03 -2.0800 0.037539 *
ENFC90
            5.9618e-02 9.4010e-03 6.3417 2.320e-10 ***
ENF3
```

La résolution du problème d'autocorrélation des résidus

- Deux possibilités:
 - 1) Corriger la matrice de variance covariance.
 - Estimateur sandwich de Huber-White / (robust standard errors):
 - On corrige les écart-types pour tenir compte de l'hétéroscédasticité (mais pas de l'autocorrélation)
 - Correction par cluster (robust clustered standard errors)
 - On corrige à la fois l'hétéroscédasticité des écarts-types et de leur autocorrélation par cluster. Les clusters sont ici les individus.
 - 2) Modéliser l'erreur comme la combinaison d'une erreur individuelle fixe et d'une erreur temporelle
 - Modèle à effet aléatoire (random effect)

Exemple: Correction Huber-White

t test of coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.2927e+00 9.0911e-02 36.2191 < 2.2e-16 ***
           2.2499e-02 1.2299e-03 18.2938 < 2.2e-16 ***
exp
           -2.4025e-04 4.5551e-05 -5.2744 1.345e-07 ***
exp2
           4.1935e-02 5.1567e-03 8.1322 4.454e-16 ***
age
           -4.5960e-04 6.8278e-05 -6.7313 1.726e-11 ***
age2
           -2.4164e-01 1.2148e-02 -19.8913 < 2.2e-16 ***
DDIPL3
DDIPL4
           -4.4744e-01 1.2162e-02 -36.7911 < 2.2e-16 ***
           -6.1445e-01 1.1590e-02 -53.0136 < 2.2e-16 ***
DDIPL5
           -6.0920e-01 1.3211e-02 -46.1112 < 2.2e-16 ***
DDIPL6
           -7.5569e-01 1.2860e-02 -58.7645 < 2.2e-16 ***
DDIPL7
M2
           -1.8977e-02 7.0873e-03 -2.6775 0.007423 **
                       1.9818e-02 -1.8309 0.067126 .
M3
           -3.6286e-02
            3.3106e-02
                       1.0064e-02 3.2894 0.001006 **
M4
           -6.9214e-03 8.1427e-03 -0.8500 0.395333
COHAB2
immig
           4.1933e-03 1.0030e-02 0.4181 0.675884
           -6.0055e-03 2.8971e-03 -2.0729 0.038190 *
ENFC90
ENF3
            5.9618e-02 8.4453e-03 7.0593 1.726e-12 ***
```

Correction RCSE

```
Estimate Std. Error t value Pr(>|t|)
           3.2927e+00 1.1151e-01 29.5286 < 2.2e-16 ***
(Intercept)
            2.2499e-02 1.5209e-03 14.7935 < 2.2e-16 ***
exp
           -2.4025e-04 5.5986e-05 -4.2913 1.785e-05 ***
exp2
           4.1935e-02 6.3389e-03 6.6155 3.795e-11 ***
age
          -4.5960e-04 8.3578e-05 -5.4990 3.865e-08 ***
age2
DDIPL3
           -2.4164e-01 1.5346e-02 -15.7457 < 2.2e-16 ***
          -4.4744e-01 1.5501e-02 -28.8655 < 2.2e-16 ***
DDIPL4
          -6.1445e-01 1.4721e-02 -41.7409 < 2.2e-16 ***
DDIPL5
DDIPL6
           -6.0920e-01 1.6992e-02 -35.8512 < 2.2e-16 ***
DDIPL7
           -7.5569e-01 1.6121e-02 -46.8756 < 2.2e-16 ***
M2
           -1.8977e-02 8.7445e-03 -2.1701 0.030009 *
M3
           -3.6286e-02 2.5739e-02 -1.4098 0.158618
M4
            3.3106e-02 1.2374e-02 2.6755 0.007469 **
COHAB2
           -6.9214e-03 9.8285e-03 -0.7042 0.481306
immiq
           4.1933e-03 1.2802e-02 0.3275 0.743262
ENFC90
           -6.0055e-03 3.6612e-03 -1.6403 0.100957
            5.9618e-02 9.4602e-03 6.3020 2.998e-10 ***
ENF3
```

Utilisation de coeftest

• Avec le package plm poo<-plm(lsalhor~exp+exp2+age+age2+DDIPL+M +COHAB+immig+ENFC90+ENF3, data=d,index=c("ident","an"),model="pooling") Coefficients normaux library("lmtest") coeftest(poo) summary(poo) - Erreurs-types robustes clusterisées : coeftest(poo, vcov=function(x) vcovHC(x, cluster="group",

type="HC1"))

Erreurs types robustes clusterisés avec R

• Fonction clx:

Utilisation de clx

• Estimation des MCO

```
reg<-lm(lsalhor~exp+exp2+age+age2+DDIPL+M
     +COHAB+immig+ENFC90+ENF3, data=d)
summary(reg)</pre>
```

• Méthode Huber White de correction de l'hétéroscédasticité seulement

```
white <- clx(reg, 1, 1:length(reg$fitted.values))
white</pre>
```

• Erreurs-types robustes clusterisées :

```
rcse <- clx(reg, 1, d$ident)
rcse</pre>
```

(Attention, l'identifiant et l'estimation doivent avoir la même taille **>** pas de non réponse ou suppression préalable des non-réponses)

Fonction pour erreurs-types robustes clusterisés selon deux groupes de clusterisation

```
mclx <-
function(fm, dfcw, cluster1, cluster2){
library(sandwich)
 library(lmtest)
 cluster12 = paste(cluster1,cluster2,
   sep="")
 M1 <- length(unique(cluster1))</pre>
 M2 <- length(unique(cluster2))</pre>
 M12 <- length(unique(cluster12))</pre>
 N <- length(cluster1)</pre>
 K <- fm$rank</pre>
 dfc1 <- (M1/(M1-1))*((N-1)/(N-K))
 dfc2 <- (M2/(M2-1))*((N-1)/(N-K))
 dfc12 \leftarrow (M12/(M12-1))*((N-1)/(N-K))
 u1 <- apply(estfun(fm), 2,</pre>
 function(x) tapply(x, cluster1, sum))
```

```
u2 <- apply(estfun(fm), 2,
  function(x) tapply(x, cluster2, sum))
u12 <- apply(estfun(fm), 2,
  function(x) tapply(x, cluster12, sum))
vc1 <- dfc1*sandwich(fm,
    meat=crossprod(u1)/N )
vc2 <- dfc2*sandwich(fm,
    meat=crossprod(u2)/N )
vc12 <- dfc12*sandwich(fm,
    meat=crossprod(u12)/N)
vcovMCL <- (vc1 + vc2 - vc12)*dfcw
coeftest(fm, vcovMCL)}</pre>
```

• Utilisation:

mclx(fm,1, test\$firmid, test\$year)

Les modèles à effet aléatoire

• On estime le modèle suivant

$$y_{it} = \beta_0 + \beta_1 x_{1it} + ... + \beta_k x_{kit} + \alpha_i + e_{it}$$

Il y a deux termes aléatoires a_i et e_{it} avec pour chacun

- Il y a deux termes aléatoires a_i et e_{it} avec pour chacun une loi de distribution propre.
- La technique d'estimation est celle des moindres carrés généralisés (lent!).
- L'avantage de ce modèle est de permettre une décomposition de la variance des résidus en un facteur individuel [i.e. between] (& le cas échéant un facteur temporel) et un facteur idiosyncrasique.

Mise en œuvre : package plm

```
install.packages("plm")
library(plm)
regre<-plm(lsalhor~exp+exp2+age
    +age2+DDIPL+M+COHAB+immig+ENFC90+ENF
    3,data=d, index=c("ident", "an"),
    model="random")
summary(regre)</pre>
```

```
Oneway (individual) effect Random Effect Model
  (Swamy-Arora's transformation)
Unbalanced Panel: n=10679, T=1-3, N=20450
Effects:
                var std.dev share
idiosyncratic 0.05070 0.22516 0.349
           0.09448 0.30737 0.651
individual
theta:
  Min. 1st Ou. Median Mean 3rd Ou.
                                        Max.
 0.4091 0.5401 0.6105 0.5480 0.6105 0.6105
Coefficients:
              Estimate Std. Error t-value Pr(>|t|)
(Intercept) 3.2800e+00 1.0037e-01 32.6794 < 2.2e-16 ***
           2.0001e-02 1.3143e-03 15.2174 < 2.2e-16 ***
exp
exp2
        -2.5480e-04 5.0115e-05 -5.0843 3.722e-07 ***
        4.0225e-02 5.5856e-03 7.2016 6.160e-13 ***
age
age2 -4.0717e-04 7.3658e-05 -5.5278 3.282e-08 ***
DDIPL3
       -2.2299e-01 1.3153e-02 -16.9530 < 2.2e-16 ***
DDIPL4
       -4.1766e-01 1.3313e-02 -31.3718 < 2.2e-16 ***
DDIPL5
           -5.8472e-01 1.2327e-02 -47.4347 < 2.2e-16 ***
DDIPL6
          -5.8619e-01 1.6221e-02 -36.1379 < 2.2e-16 ***
           -7.2696e-01 1.3231e-02 -54.9446 < 2.2e-16 ***
DDIPL7
           -1.4219e-02 9.0952e-03 -1.5633 0.1179903
M2
           -3.4741e-02 3.0758e-02 -1.1295 0.2586942
М3
           2.2202e-02 1.2503e-02 1.7757 0.0757952 .
M4
COHAB2
           -1.6166e-03 9.3976e-03 -0.1720 0.8634242
immig
       1.8494e-03 1.2368e-02 0.1495 0.8811359
           -9.0373e-03 3.4866e-03 -2.5920 0.0095494 **
ENFC90
            3.2673e-02 8.6185e-03 3.7911 0.0001504 ***
ENF3
```

Comparaison de la correction de l'autocorrélation des erreurs

 Le modèle pooling suppose la matrice de variance covariance suivante : même variance pour tous les résidus, pas d'autocorrélation des résidus

$$u'u = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

• La correction de White Huber permet une hétérogénéité des variances des résidus

$$u'u = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3^2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

correction de l'autocorrélation des

erreurs

• Ecart-types robustes clusterisés (RCSE).

σ_{j1}^2	$\rho_{j12}\sigma_{j1}{}^2\sigma_{j2}{}^2$
$\rho_{j12}\sigma_{j1}{}^2\sigma_{j2}{}^2$	σ_{j2}^2

		i1	i2	i3	j1	j2	j3		n1	n2	n3
	i1	σ_{i1}^2	$\rho_{i12}\sigma_{i1}{}^2\sigma_{i2}{}^2$	$\rho_{i13}\sigma_{i1}{}^2\sigma_{i3}{}^2$	0	0	0 /	<i>_</i>	0	0	0
	i2	$\rho_{i12}\sigma_{i1}{}^2\sigma_{i2}{}^2$	σ_{i2}^{2}	$\rho_{i23}\sigma_{i2}{}^2\sigma_{i3}{}^2$	0	0	0		0	0	0
	i3	$\rho_{i13}\sigma_{i1}{}^2\sigma_{i3}{}^2$	$\rho_{i23}\sigma_{i2}{}^2\sigma_{i3}{}^2$	σ_{i3}^2	0	0	0		0	0	0
	j1	0	0	0	σ_{j1}^2	$\rho_{j12}\sigma_{j1}^{2}\overline{\sigma}_{j2}^{2}$	$\rho_{j13}\sigma_{j1}{}^2\sigma_{j3}{}^2$		0	0	0
u'u=	j2	0	0	0	$\rho_{j12}\sigma_{j1}{}^2\sigma_{j2}{}^2$	σ_{j2}^{2}	$\rho_{j23}\sigma_{j2}{}^2\sigma_{j3}{}^2$		0	0	0
	j3	0	0	0	$\rho_{j13}\sigma_{j1}{}^2\sigma_{j3}{}^2$	$\rho_{j23}\sigma_{j2}{}^{2}\sigma_{j3}{}^{2}$	$\sigma_{j3}{}^{2}$		0	0	0
	n1	0	0	0	0	0	0		σ_{n1}^2	$\rho_{n12}\sigma_{n1}{}^2\sigma_{n2}{}^2$	$\rho_{n13}\sigma_{n1}{}^2\sigma_{n3}{}^2$
	n2	0	0	0	0	0	0		$\rho_{n12}\sigma_{n1}{}^2\sigma_{n2}{}^2$	σ_{n2}^{2}	$\rho_{n23}\sigma_{n2}{}^2\sigma_{n3}{}^2$
	n3	0	0	0	0	0	0		$\rho_{n13}\sigma_{n1}{}^2\sigma_{n3}{}^2$	$\rho_{n23}\sigma_{n2}{}^2\sigma_{n3}{}^2$	σ_{n3}^2

• Effets aléatoires

Éliminer la variance interindividuelle et se concentrer sur la variance intraindividuelle

... et régler le problème de l'hétérogénéité inobservée (constante)

Le problème d'hétérogénéité inobservée

• Revenons sur les bases du modèle à effet aléatoire :

$$y_{it} = \beta_0 + \beta_1 x_{1it} + ... + \beta_k x_{kit} + \alpha_i + e_{it}$$

- Si $cov(x_{kit}, a_i) = 0$, alors pas de pb.
- Mais l'erreur individuelle *a_i* peut-être vue comme le produit de l'ensemble des variables inobservables invariantes dans le temps... et il n'est pas impossible que ces variables inobservables soient corrélées avec les observables.
- Plutôt que de modéliser les a_i par un effet aléatoire, on peut les modéliser par un effet fixe.

Solution 1: Effets fixes (fixed effects)

- Moindre carré à variables dichotomiques : Least square dummy variables LSDV
 - Application soit à introduire une variable dichotomique a_i par individu

$$y_{it} = \beta_0 + \beta_1 x_{1it} + ... + \beta_k x_{kit} + a_i + e_{it}$$

- Possible uniquement si le nombre d'individus petit (<1000).
- Modèle « within »
 - modèle (sans constante) où l'on centre chaque variable (explicative ou expliquée), c'est-à-dire où l'on calcule l'écart à la moyenne de la variable pour l'individu.

$$(y_{it} - \overline{y}_i) = \beta_1(x_{1it} - \overline{x}_{1i}) + \dots + \beta_k(x_{kit} - \overline{x}_{ki}) + (e_{it} - \overline{e}_i)$$

Solution 2 : Modèles en différences premières (first difference)

• Expliquer les évolutions par les évolutions

$$(y_{it} - y_{it-1}) = \beta_0 + \beta_1(x_{1it} - x_{1it-1}) + \dots + \beta_k(x_{kit} - x_{kit-1}) + (e_{it} - e_{it-1})$$

- Le choix du retard
 - [t,t-1] → on se focalise sur des variations de court terme
 - [t,t-k]→ variations de plus long terme... mais perte d'effectifs et de puissance
- Différence entre effets fixes & différences premières
 - Panel à deux périodes : modèles équivalents
 - Panel à plusieurs périodes : effets fixes est une moyenne de variations de courtes et de longues périodes

Estimation des effets fixes

• Avec les effets fixes :

```
regfed<-lm(lsalhor~factor(ident)+exp+exp2+age+age2+DDIPL
+M+COHAB+immig+ENFC90+ENF3,data=d)
summary(regfed)
```

• Avec l'estimateur within et le package plm

• Avec l'estimateur within et le package lfe

```
library ("lfe")
fixb<-felm(lsalhor~exp+exp2+age+age2+DDIPL
+M+COHAB+immig+ENFC90+ENF3|ident, data=d)
```

```
Oneway (individual) effect Within Model
Unbalanced Panel: n=10679, T=1-3, N=20450
Coefficients:
         Estimate Std. Error t-value Pr(>|t|)
exp 5.3791e-03 2.2145e-03 2.4290 0.0151571 *
exp2 -1.8009e-04 8.2044e-05 -2.1951 0.0281822 *
age 7.9820e-02 1.5302e-02 5.2162 1.864e-07 ***
age2 -3.9801e-04 1.9414e-04 -2.0501 0.0403795 *
DDIPL3 -1.9272e-01 9.5241e-02 -2.0235 0.0430473 *
DDIPL4 -5.3933e-01 1.5849e-01 -3.4028 0.0006697 ***
DDIPL5 -1.6766e-01 1.4452e-01 -1.1601 0.2460355
DDIPL6 -3.0187e-01 1.8245e-01 -1.6545 0.0980633 .
DDIPL7 -8.3179e-02 1.8803e-01 -0.4424 0.6582296
M2 -3.6406e-03 1.6870e-02 -0.2158 0.8291488
M3 1.9240e-02 7.0249e-02 0.2739 0.7841852
M4 1.8409e-02 2.2369e-02 0.8230 0.4105527
COHAB2 1.2513e-02 2.0936e-02 0.5977 0.5500597
ENFC90 -4.1538e-03 8.5598e-03 -0.4853 0.6275027
ENF3 -8.9639e-04 1.1233e-02 -0.0798 0.9363962
Signif. codes: 0 \*** 0.001 \** 0.01 \*' 0.05 \.' 0.1 \' 1
Total Sum of Squares: 517.96 Residual Sum of Squares: 494.6
R-Squared : 0.045094 Adj. R-Squared : 0.021513
F-statistic: 30.7141 on 15 and 9756 DF, p-value: < 2.22e-16
```

Modèles en différences premières

Table 3. Coefficients from Difference Models Regressing Differences in Ln Hourly Wage Over Time on Differences in Selected Family Status Variables: Women from the NLS-YW, 1968–1988

	Model and Number of Years between Observations								
Family Status	Model 1 (1-2 Years)	Model 2 (2–4 Years)	Model 3 (3–5 Years)	Model 4 (5–9 Years)					
Married	.003	.020*	.027*	.043*					
Divorced	.028*	.032*	.033*	.061*					
One child	018	053*	061*	064*					
Two or more children	023	056*	076*	111 *					
Children not in the home	004	009*	010*	016 *					
Number of observations	21,460	18,026	15,535	11,559					

Note: Coefficients are from difference models in which the dependent variable is the difference between the natural log hourly wage for an individual in one year and the natural log hourly wage for that individual in the comparison year. The independent variables are expressed as differences as well and include actual experience, experience squared, age, age squared, education, separated, and widowed. Observations are woman-years.

^{*}p < .05 (two-tailed tests)

Estimation des modèles en différences premières

Oneway (individual) effect First-Difference Model Unbalanced Panel: n=10679, T=1-3, N=20450 Coefficients:

```
Estimate Std. Error t-value Pr(>|t|)
(intercept) 4.2842e-02 1.6268e-02 2.6336 0.008462 **
exp 5.8668e-03 2.2880e-03 2.5641 0.010358 *
exp2 -1.6842e-04 8.4906e-05 -1.9836 0.047332 *
age 4.4148e-02 2.3722e-02 1.8611 0.062761 .
age2 -4.7521e-04 2.3337e-04 -2.0363 0.041751 *
DDIPL3 -2.0762e-01 9.9803e-02 -2.0802 0.037529 *
DDIPL4 -5.8213e-01 1.6363e-01 -3.5576 0.000376 ***
DDIPL5 -1.7071e-01 1.4998e-01 -1.1382 0.255076
DDIPL6 -2.6763e-01 1.9138e-01 -1.3984 0.162024
DDIPL7 -8.8853e-02 1.8937e-01 -0.4692 0.638936
M2 4.7657e-03 1.7291e-02 0.2756 0.782843
M3 4.8832e-02 6.9279e-02 0.7049 0.480918
M4 1.6575e-02 2.2434e-02 0.7388 0.460042
COHAB2 -7.5124e-03 2.1919e-02 -0.3427 0.731805
ENFC90 -2.3596e-03 9.1098e-03 -0.2590 0.795632
ENF3 -2.6791e-03 1.1977e-02 -0.2237 0.823012
```

Signif. codes: 0 ***' 0.001 **' 0.01 *' 0.05 \'.' 0.1 \' 1

Total Sum of Squares: 958.2 Residual Sum of Squares: 955.09

R-Squared: 0.0032506 Adj. R-Squared: 0.0032453

F-statistic: 2.12087 on 15 and 9755 DF, p-value: 0.0068917

Effets fixes et DP. Remarques

- Double effet : réduction du problème d'autocorrélation et suppression du problème d'hétérogénéité inobservée invariante dans le temps.
- Interprétation : Les évolutions expliquent les évolutions.
- LSDV : Variables dichotomiques en très grand nombre (quand on les estime) => On n'imprime généralement pas les résultats !
- Disparition de la constante en effets fixes mais pas en différences premières

Effets fixes et différences premières. Remarques

- Disparition de toutes les variables constantes dans le temps
 effet contenu dans les variables dichotomiques individuelles.
 - Ex: immigration
 - Parfois les variables présumées constantes se maintiennent en raison de points aberrants:
 - incohérence de déclaration d'une vague à l'autre
 - Changement de sexe
 - Diplômes tardifs...
- On peut les introduire malgré tout en les croisant avec une variable temporelle. L'interprétation reste en termes d'évolution.)

```
ENFC90 -4.311e-03 8.557e-03 -0.504 0.614427
ENF3 -9.663e-04 1.123e-02 -0.086 0.931418
immig NA NA NA NA NA
an 6.513e-02 2.346e-02 2.777 0.005497 **
immig:an 1.358e-02 8.548e-03 1.589 0.112169
```

Les deux effets fixes : groupes et temps

- On peut introduire un effet fixe par unité de groupe (ex. Individu, pays, etc.) et un effet fixe par unité de temps
- \(\rightarrow\) la conjoncture commune capturée par l'effet fixe temps
 - Tout avec des variables dichotomiques :

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \dots + \beta_k x_{kit} + a_i + d_t + e_{it}$$

- Within + variables dichotomiques de temps

$$(y_{it} - \overline{y}_i) = \beta_1(x_{1it} - \overline{x}_{1i}) + ... + \beta_k(x_{kit} - \overline{x}_{ki}) + d_t + (e_{it} - \overline{e}_i)$$

- FD + variables dichotomiques de temps

$$(y_{it} - y_{it-1}) = \beta_0 + \beta_1(x_{1it} - x_{1it-1}) + \dots + \beta_k(x_{kit} - x_{kit-1}) + d_t + (e_{it} - e_{it-1})$$

- Two -ways within

$$(y_{it} - \overline{y}_i - \overline{y}_t + \overline{y}) = \beta_1 (x_{1it} - \overline{x}_{1i} - \overline{x}_{1t} + \overline{x}_1) + ... + \beta_k (x_{kit} - \overline{x}_{ki} - \overline{x}_{kt} + \overline{x}_k) + (e_{it} - \overline{e}_i - \overline{e}_t - \overline{e})$$

$$w2 < -plm(lsalhor \sim exp + ... + ENF3, index = c("ident", "an"), model = "within", effect = "twoways", data = d)$$

Se concentrer sur la variance interindividuelle et éliminer la variance intra-individuelle

Étudier la variance interindividuelle

- Le modèle à effets fixes ou within permet d'étudier les évolutions intra-individuelles
- La régression homogène (pooled) ou celles à effets aléatoires sont deux manières d'analyser une combinaison de variations intra-individuelles et interindividuelles
- Peut-on étudier les variations strictement interindividuelles ?
- Oui, de deux façons
 - Régression toute simple sur une seule période.
 - Régression between.

Régression between

• La régression between est une régression où l'on explique la moyenne par individu de la variable expliquée par les moyennes par individu des variables explicatives.

$$\overline{y}_i = \beta_1 \overline{x}_{1i} + \dots + \beta_k \overline{x}_{ki} + \overline{u}_i$$

- Cette régression informe sur la variation interindividuelle.
- Elle peut être considérée comme la bonne régression :
 - Si la variation intra-individuelle dans le temps est négligeable
 - Si les individus qui connaissent des variations intra-individuelles sont non représentatifs.

Oneway (individual) effect Between Model Unbalanced Panel: n=10679, T=1-3, N=20450 Coefficients:

```
Estimate Std. Error t-value Pr(>|t|)
(Intercept) 3.2669e+00 1.1063e-01 29.5291 < 2.2e-16 ***
          2.3898e-02 1.6630e-03 14.3701 < 2.2e-16 ***
exp
         -2.5506e-04 6.4273e-05 -3.9684 7.283e-05 ***
exp2
          4.2127e-02 6.1914e-03 6.8041 1.071e-11 ***
age
age2 -4.7735e-04 8.1889e-05 -5.8292 5.731e-09 ***
          -2.2276e-01 1.3685e-02 -16.2771 < 2.2e-16 ***
DDIPL3
DDIPL4 -4.1411e-01 1.3801e-02 -30.0071 < 2.2e-16 ***
          -5.8193e-01 1.2809e-02 -45.4316 < 2.2e-16 ***
DDIPL5
DDIPL6 -5.7433e-01 1.6899e-02 -33.9855 < 2.2e-16 ***
DDIPL7 -7.0636e-01 1.3721e-02 -51.4807 < 2.2e-16 ***
          -1.9630e-02 1.1126e-02 -1.7644 0.07770 .
M2
М3
          -2.9286e-02 3.5068e-02 -0.8351 0.40367
          3.4091e-02 1.5360e-02 2.2195 0.02647 *
M4
COHAB2 -4.9084e-03 1.1094e-02 -0.4424 0.65817
immig 1.5252e-02 1.2769e-02 1.1945 0.23232
          -7.5518e-03 3.9287e-03 -1.9222 0.05461 .
ENFC90
ENF3
          6.7982e-02 1.3868e-02 4.9022 9.618e-07 ***
Signif. codes: 0 \***' 0.001 \**' 0.01 \*' 0.05 \.' 0.1 \' 1
Total Sum of Squares: 2305.4 Residual Sum of Squares: 1544.8
R-Squared : 0.32995 Adj. R-Squared : 0.32942
F-statistic: 328.133 on 16 and 10662 DF, p-value: < 2.22e-16
```

Relations entre les méthodes et comparaison de leur pertinence

Modèle aléatoire entre effet fixe et modèle between

- Il existe une relation mathématique entre la régression à effets aléatoires, la régression à effets fixes et la régression sur les moyennes:
- $\beta_{\text{Random}} = \Delta \beta_{\text{Within}} + (I-\Delta) \beta_{\text{Between}}$
 - Où ∆ est une matrice compliquée...

Modèle aléatoire entre pooling et effets fixes

• Un paramètre λ mesure aussi la proximité entre le modèle à effets fixes et la régression *pooling*.

$$- \lambda = 1 - [\psi]^{0.5} \qquad \qquad Ou \qquad \psi = \sigma_e^2 / (\sigma_e^2 + T. \sigma_a^2)$$

- Quand Ψ tend vers 0 (et donc λ vers 1) on tend vers l'estimateur à effet fixe (*within*) fondé uniquement sur la variance intra-individuelle.
- Quand Ψ tend vers 1 (et donc λ vers 0) on tend vers l'estimateur de *pooling* qui tient compte de la variance totale c'est à dire à la fois de la variance intra individuelle et inter individuelle
- L'introduction d'effets individuels aléatoires permet d'obtenir une spécification intermédiaire entre le modèle sans effet individuel et le modèle avec effets fixes. L'hypothèse d'une distribution commune des effets individuels permet une d'adopter une structure ni totalement homogène ni totalement hétérogène. Les effets individuels ont en commun une distribution identique.

Une combinaison du pooling et des effets fixes

Le modèle aléatoire

$$y_{it} = \alpha + \beta X_{it} + \mu_i + e_{it}$$

- Ainsi, à la place de la transformation "within", on enlève <u>une partie</u> de la variation within :
 - GLS Estimator: "Quasi time-demeaned data"

$$y_{it} - \lambda \overline{y} = \beta (X_{it} - \lambda \overline{X}) + (e_{it} - \lambda e_i)$$

• où:

$$\lambda = 1 - \sqrt{\sigma_v^2 / (\sigma_v^2 + T\sigma_\mu^2)}$$

Exemple

Between Pooling Within Enfc90 -0.0076· -0.0060* -0.0042 Enf3 0.0680*** 0.0596*** -0.0009

Random

-0.0090**

0.0327***

L'économétrie des panels réécrite avec des statistiques descriptives

		Pas d' enfant <3	Au moins 1 enfant	
Équivalent régression	Variable	an	<3 an	Différence
Between	Interindividuel	3.85	3.89	+4.6%**
Pooling	Interindividuel et intraindividuel	3.87	3.91	+3.8%***
Random	Interindividuel et intraindividuel (++)	3.85	3.86	+1.1
Within	Intraindividuel		,	-0.5%

		Sortie		Entrée	
		enfant	Pas de	enfant	
Équivalent régression	Variable	<3	changement	<3	Différence
First Difference	Interindividuel	0.068	0.051	0.062	-0.3%

Les effets aléatoires

- Le modèle à effet aléatoire est un hybride: Un modèle intermédiaire entre MCO et effets fixes
 - Si T est important, cela ressemble à des effets fixes
 - Si l'effet unité a une faible variance (peu important), modèle à effet aléatoire ressemble aux MCO
 - Si l'effet unité est important (unités très divergentes), résultats ressemblent aux effets fixes
 - Modèle à effets aléatoires avec Moindres carrés généralisés est plutôt pour des enquêtes à grand N
 - Ses propriétés pour les petits N et grand T ne sont pas très bien étudiés
 - Beck et Katz conseillent de ne pas les utiliser
 - Et, si on doit les utiliser, ils recommandent le max. de vraisemblance

Effets aléatoires

- Quand utiliser les effets aléatoires?
 - Quand on pense que les effets unitaires inobservés ne sont pas corrélés avec les variables explicatives X
 - Soit pour des raisons théoriques ...
 - Ou, parce que vous avez des bons contrôles
 - Ex: des variables suprarégionales pour les pays...
 - Vous vous intéressez aux variations constantes dans le temps
 - Quand vous vous intéressez aux différences entre unité
 - Quand il y a très peu de variations
 - Quand le test d'Hausman indique que les résultats sont similaires

Effets fixes ou effets aléatoires?

- Comment choisir ?
- Test de Hausman : on regarde si l'estimation par effets fixes produit des résultats significativement différents d'un modèle à effets aléatoires
- Si le test de Hausman est significatif, on privilégiera les effets fixes.
- Si le test de Hausman n'est pas significatif, on privilégiera les effets aléatoires.

Mise en œuvre

```
regfe<-plm(lsalhor~exp+exp2+age
  +age2+DDIPL+M+COHAB+immig+ENFC90+ENF3,data=d,
  index=c("ident", "an"), model="within")
regre<-plm(lsalhor~exp+exp2+age
  +age2+DDIPL+M+COHAB+immig+ENFC90+ENF3,data=d,
  index=c("ident", "an"), model="random")
phtest(regfe, regre)
      Hausman Test
data: lsalhor ~ exp + exp2 + age + age2 + DDIPL + M
  + COHAB + immig + ...
chisq = 488.0588, df = 15, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

Econométrie des panels avec Stata

Préparation du panel xtset ident an Effets aléatoires xtreg lsalhor exp exp2 age age2 ENFC90 ENF3, re Effets fixes xtreg lsalhor exp exp2 age age2 ENFC90 ENF3, fe Between xtreg lsalhor exp exp2 age age2 ENFC90 ENF3, be Test d'Hausman effets fixes versus effets aléatoires quietly xtreg lsalhor exp exp2 age age2 ENFC90 ENF3, fe estimates store fixed quietly xtreg lsalhor exp exp2 age age2 ENFC90 ENF3, re estimates store random

hausman fixed random

Econométrie des panels avec SAS

```
D'abord trier les données
 Proc sort data=a;
 By ident an;
 Run;
Sas 9.2 et ultérieur : Proc panel
 Proc panel data=a ranone rantwo btwng pooled fixone fixtwo;
 Id ident an;
 Model y=x1 x2;
 Run;
 - Ranone : effet aléatoire individuel. Rantwo : effet aléatoire individuel et temporel. Btwng : between.
     Pooled: Pooled. Fixone: effets fixe individuels. Fixtwo: effets fixes individuels et temporels.
Within, avec GLM. Evite l'estimation des paramètres individuels!
 Proc qlm data=a;
 Absorb ident;
 Model y=x1 x2;
 Run;
```

Les effets fixes hors données de panel

- Effets fixes => établissement, classes, etc.
- Restreindre les espaces de comparaison.
 - localisme et productivité
- Modèles de la différence de différence
- Attention à la restriction des données : biais de sélection à exclure les observations pour lesquelles rien ne change

Variations intraindiv. de la		
variable explicative	Oui	Non
de la variable expliquée		
Oui	Données utiles	Données indirectement utiles
Non	Données indirectement utiles	Données non-utiles

Limites des effets fixes

- On estime les effets sur les individus qui connaissent des évolutions
 - Si y_i ne change pas en fonction de t, alors y_i capturé par l'effet fixe a_i .
 - Ces évolutions, surtout sur des variables explicatives qualitatives, peuvent être estimées sur des individus rares et très singuliers : biais de sélection
 - Les variables explicatives du changement peuvent être aussi très singulières
 - Ex extrême: changement de sexe.
 - Cela revient à considérer que la différence d'outcome entre les sexes est estimée par l'évolution de l'outcome pour ceux qui changent de sexe
 - Ex classique: promotion non-cadre → cadre
 - Les effets fixes reviennent à considérer que les évolutions estimées sur les gens qui changent sont représentatives des différence d'état entre gens qui changent pas.

Limites des effets fixes

- Les effets fixes permettent bien de corriger l'hétérogénéité inobservée! Mais une partie seulement
 - L'hétérogénéité inobservée invariante dans le temps....
- Mais... transforment une régression en une régression en évolution...
 - Où un facteur d'évolution corrélé aux autres peut être inobservé et conduire à biaiser les autres
 - Où il peut y avoir simultanéité des évolutions de la variable indépendante et des variables indépendantes
 - → variables instrumentales (si on en trouve)
- Pas de correction de l'hétérogénéité inobservée qui varie dans le temps...

Limite des effets fixes

- Ontologie essentialiste
 - Effet individuel : Fonds inchangé, permanent, toujours le même
- Les évolutions expliquent les évolutions, mais le fixe n'explique pas le changement (sauf spécification contraire)

Extension logistique

Extension des panels à des modèles non linéaires

- Probit (effets aléatoires), Logistique (effets fixes, effets aléatoires)
 - NB: Modèles probit déconseillés avec des effets fixes
- Package pglm :
 - binomial models (logit and probit) count models (poisson and negbin)
 and ordered models (logit and probit)
 - Syntaxe (exemple):

```
ralog <-pglm(ACT3~exp+exp2+age+age2+COHAB+ENFC90 +ENF3, index="ident", data=d, family=binomial, model="random")
```

- Actuellement, la méthode within et between ne fonctionne pas pour les modèles logit et probit avec pglm.
- Package survival
 - clogit

```
felog<-clogit(ACT3~exp+exp2+age+age2+COHAB+ENFC90+ENF3+strata(ident), data=d)
```

Logit en Panel sous STATA

• Effets fixes

```
xtlogit inact age age2 ENFC90 ENF3, re estimates store random
```

• Effets aléatoires

```
xtlogit inact age age2 ENFC90 ENF3, fe
estimates store fixed
```

• Hausman test

hausman fixed random

Logit en Panel sous SAS

• Effets fixes

```
Proc logistic data=a;
Strata ident;
Model y=x1 x2;
Run;
```

- Effets aléatoires
 - Proc glimmix

Safi, Rathelot Le white flight existe-t-il en France?

Share of dropouts	.969***	.007	1.042	.026
Log of total population	.870***	.014	.775***	.026
For Municipalities > 10,000 Inhabitants				
Share of immigrants	1.127***	.015	1.162***	.030
Share of co-ethnics (for immigrants)			.812***	.022
For Municipalities < 10,000 Inhabitants				
Share of immigrants	1.032***	.009	.996	.026
Share of co-ethnics (for immigrants)			.896***	.020

- En apparence oui.
 - Les natifs déménagent d'autant plus que les immigrés sont nombreux
- Mais déménagent-ils à cause des immigrés ou à cause de caractéristiques inobservées corrélées à la présence des immigrés?

Introduction d'un effet fixe commune

	(1)		(2)		(3)	
	Coef.	SE	Coef.	SE	Coef.	SE
Communes > 10,000 Inhabitants Natives						
Share of immigrants	.040	.034	004	.046	.048	.048
Immigrants						
Share of immigrants	.005	.040	.149	.138	007	.076
Share of co-ethnics	234***	.023	236***	.030	292***	.029
Communes < 10,000 Inhabitants Natives						
Share of immigrants Immigrants	.018	.022				
Share of immigrants	042	.038				
Share of co-ethnics	112***	.024				
Commune FE	Yes		Yes, interacted with immigrant dummy		Yes	
Individual Heterogeneity	No		No		Random Effects	
Communes' Sample	All		> 10,000 inhabitants		> 10,000 inhabitants	
N	583,	266	287,844		288,442	
Pseudo R-Sq	.1		.05		.05	

- Plus de white flight
- Une tendance à l'entresoi des communau tés immigrées

En même temps dans l'annexe

Table S3. Specifications Controlling for Individual Heterogeneity

	Individual Heterogeneity					
	(4)	(5)				
	Coef.	SE	Coef.	SE		
Communes > 10,000 Inhabitants						
Natives						
Share of immigrants	.16***	.02	.20***	.01		
Immigrants						
Share of immigrants	.02	.07	.02	.03		
Share of co-ethnics	15*	.07	31***	.03		
Communes < 10,000 Inhabitants						
Natives						
Share of immigrants	.02	.02	.06***	.01		
Immigrants						
Share of immigrants	.09	.08	.01	.04		
Share of co-ethnics	03	.06	15***	.03		
Commune FE	No		No			
Individual FE/RE	FE		RE			
Communes' Sample	ALL		ALL			
N	142,65	2	611,486			
Pseudo R-Sq	.14					

^{*}p < .05; **p < .01; ***p < .001 (two-tailed z-tests)..

Explications

- "The fixed-effect modeling leads to a significant loss in data: only observations for which a change in geographic mobility is observed across our periods contribute to the estimation (N = 142,652)."
- "Conversely, whereas random-effects models can be estimated on the whole sample, they rely on the questionable assumption that no correlation exists between individual heterogeneity and the control variables."
- "Accounting for individual heterogeneity thus does not modify the results of simple regression analyses. This supports our conclusion regarding the importance of controlling for local unobserved characteristics when studying geographic mobility."
 - NB: on suit les enquêtés ici sur plusieurs communes successives

Références

Waldfogel Jane, 1997, "The effect of Children on woman's wage", ASR

Rathelot, Roland et Safi, Mirna. Local Ethnic Composition and Natives' and Immigrants' Geographic Mobility in France, 1982–1999. *American Sociological* Review, 2013, p. 0003122413514750.