Lecture 4. Panel Regressions

Olivier Godechot

Sciences Po

M2. Sociology Master

Panel: example

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

- What's producing children wage penalty for female workers?
  - Unobserved heterogeneity: No
  - Part-time work: Yes

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

	Pooled OLS Models			Fixed-Effects Models		
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 (Definition)

- We speak of panel data when we have several observations at different dates for the same individual.
- It's common to oppose panel data (or longitudinal data) and cross-sectional data
- Panel's advantages
  - Individual evolutions
  - Parameter estimation not only based on inter-individual differences but also on intra-individual differences
  - Distinction of age and generation effects
  - Possibility to introduce lag effects.
  - Enable to establish more convincing causality relations

A few French panels commonly used

- French Labor Force Survey *(Emploi)*
  - Housing panel
  - Before 2002 three year rotating panel with one third renewal every year (three interrogations)
  - Starting 2003, 18 months rotating panel with one sixth renewal every trimester (six interrogations)
- Panel SILC (2005-.) / SRCV
  - 4-year rotating panel (9 in France) with interrogation every year
  - Panel of individuals and of household (complexity of use)
  - Thematic inquiry each year
  - Before: European Panel of households (1994-2001)
- Univers culturels des enfants et des adolescents - 2002-2008 / *Cultural universe of children and teenagers*
- Panel des élèves du second degré – *Panel of students in the secondary school level*
- Échantillon démographique permanent – *Permanent demographic census*
- Panel des DADS – *DADS panel*

Data organization (1)

- "Long format". Observations for the same individual are above one another.
- Necessary for many regression software (notably for R plm package)
- Common to put individual and time identifiers in the two first columns

	ident	an	S	SALRED	HH	salhor	exp	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

Data organization (2)

- "Wide format"

	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 of the number of columns especially if time dimension is important
- Not very practical for some manipulations
- Could be easier for others

Data organization (3)

- Balanced panel
  - => same number of observations for each individual.
- Unbalanced panel
  - Missing data, attrition, rotating panel, etc.
  - Selection bias due to attrition phenomena
- Econometric results first established on balanced panel and further extended to unbalanced ones.

Panels as problems and solution

Panel econometrics:
Problems and solutions

Panel econometrics renders visible but also offers solutions to two classical problems

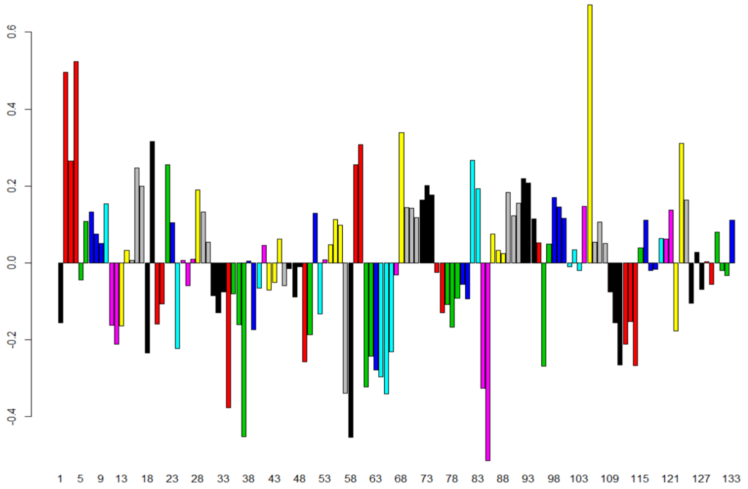
1. Autocorrelation of residuals

2. Unobserved heterogeneity

Autocorrelation of residuals

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

- $\mathrm{Cov}(u_{it}, u_{it+1}) = 0$?
- If I'm under-payed at date t, I will probably be the similarly under-payed at date t+1



Unobserved heterogeneity

- The unobserved heterogeneity problem (or that of confounding variables) can be raised in panel regression as in classical OLS

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

If $u_{it} = c.unobs + e$ and $cov(x_{ik}, unobs) \neq 0$ … then bias !

- The residual u_{it} can be rewritten as follows: $u_{it} = a_i + e_{it}$, as the sum of an individual constant error a_i , and e_{it} temporary error.
- a_i can be seen as the product of time invariant unobserved variables

Unobserved heterogeneity (2)

- Question: Can we say that our true model with this individual constant error a_i is independent from the independent variables?
- Thus, is $cov(a_i , x_k) = 0$?
  - If yes, either ***pooling regressions*** or ***random effects*** ones will the best
  - If no, then OLS $cov(u_i , x_k) = 0$ hypothesis is not respected
  - OLS empirical regression will not enable to estimate the β
  - But we can solve this problem with ***fixed effects*** or ***first differences*** models

Combining within-individual and between-individual variances

… and solving the auto-correlation of residuals problem

Pooled regression

- We can estimate the following equation with OLS.

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

- « *Pooled* or *pooling* regression » or homogenous model
  - Either simply with lm model

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

  - Or with plm package

```
poo<-plm(lsalhor~exp+exp2+age+age2+DDIPL+M
   +COHAB+immig+ENFC90+ENF3,
   data=d,index=c("ident","an"),model="pooling")
summary(poo)
```

Example wage and kids

```
    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 ***
exp          2.2499e-02  1.1765e-03  19.1236 < 2.2e-16 ***
exp2        -2.4025e-04  4.4576e-05  -5.3898 7.132e-08 ***
age          4.1935e-02  4.7122e-03   8.8993 < 2.2e-16 ***
age2        -4.5960e-04  6.1953e-05  -7.4185 1.231e-13 ***
DDIPL3      -2.4164e-01  1.0152e-02 -23.8018 < 2.2e-16 ***
DDIPL4      -4.4744e-01  1.0222e-02 -43.7716 < 2.2e-16 ***
DDIPL5      -6.1445e-01  9.4473e-03 -65.0394 < 2.2e-16 ***
DDIPL6      -6.0920e-01  1.2280e-02 -49.6070 < 2.2e-16 ***
DDIPL7      -7.5569e-01  1.0135e-02 -74.5631 < 2.2e-16 ***
M2          -1.8977e-02  8.0101e-03  -2.3691  0.017842 *
M3          -3.6286e-02  2.5193e-02  -1.4403  0.149789
M4           3.3106e-02  1.0946e-02   3.0246  0.002493 **
COHAB2      -6.9214e-03  8.2204e-03  -0.8420  0.399814
immig        4.1933e-03  9.4416e-03   0.4441  0.656952
ENFC90      -6.0055e-03  2.8873e-03  -2.0800  0.037539 *
ENF3         5.9618e-02  9.4010e-03   6.3417 2.320e-10 ***
```

Solving the residual autocorrelation problem

- Two possibilities:
  - 1) Robust clustered standard errors models
    - Correct the residuals' variance-covariance matrix.
      - We correct parameters' standard errors to account both for residuals heteroscedasticity and their within-cluster autocorrelation.
      - Clusters here are generally individuals.
      - Extension of Huber-White Sandwich robust standard errors estimator
        - » Which corrects the standard-errors to account for residuals heteroscedasticity (but not for autocorrelation
  - 2) Random effect models
    - Modeling error as a combination of a time invariant individual error and a temporary error

Example: Huber-White Correction

```
t test of coefficients:

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

Robust Cluster Standard Errors

```
               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  3.2927e+00  1.1151e-01  29.5286 < 2.2e-16 ***
exp          2.2499e-02  1.5209e-03  14.7935 < 2.2e-16 ***
exp2        -2.4025e-04  5.5986e-05  -4.2913 1.785e-05 ***
age          4.1935e-02  6.3389e-03   6.6155 3.795e-11 ***
age2        -4.5960e-04  8.3578e-05  -5.4990 3.865e-08 ***
DDIPL3      -2.4164e-01  1.5346e-02 -15.7457 < 2.2e-16 ***
DDIPL4      -4.4744e-01  1.5501e-02 -28.8655 < 2.2e-16 ***
DDIPL5      -6.1445e-01  1.4721e-02 -41.7409 < 2.2e-16 ***
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
immig        4.1933e-03  1.2802e-02   0.3275  0.743262
ENFC90      -6.0055e-03  3.6612e-03  -1.6403  0.100957
ENF3         5.9618e-02  9.4602e-03   6.3020 2.998e-10 ***
```

RCSE with LFE and with PLM

- Easy implementation with LFE

```
install.packages("lfe")
library("lfe")
poo<-felm(lsalhor~exp+exp2+age+age2+DDIPL+M
    +COHAB+immig+ENFC90+ENF3|0|0|ident,
    data=d)
```

- With plm package

```
poo<-plm(lsalhor~exp+exp2+age+age2+DDIPL+M
    +COHAB+immig+ENFC90+ENF3,
    data=d,index=c("ident","an"),model="pooling")
```

  - Normal Coefficients

```
library("lmtest")
coeftest(poo)
summary(poo)
```

  - Robust cluster standard errors :

```
coeftest(poo,
  vcov=function(x) vcovHC(x, cluster="group", type="HC1"))
```

With a homemade functions with R

- With clx function:

```
clx <- function(fm, dfcw, cluster){
  library(sandwich)
  library(lmtest)
  M <- length(unique(cluster))
  N <- length(cluster)
  dfc <- (M/(M-1))*((N-1)/(N-fm$rank))
  u <- apply(estfun(fm),2,
             function(x) tapply(x, cluster, sum))
  vcovCL <- dfc*sandwich(fm, meat=crossprod(u)/N)*dfcw
  coeftest(fm, vcovCL) }
```

Use of clx

- OLS Estimation

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

- Huber White correction of heteroscedasticity only

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

- Robust cluster standard errors:

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

(Beware, ID (d$ident) and regression estimates need to have the same size ➔ no missings or you have to suppress the missings)

Tip: `clx(reg,1,factor(d$ident[-attr(reg$model,"na.action")]))`

Function for two-way robust cluster standard errors

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

- Use:

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

With Stata

- Homogenous model (pooling)

```
reg lsalhor exp exp2 age age2 i.diplome i.matrim i.couple immig i.an ENFC90 ENF3
```

- Homogenous model (pooling) and Robust standard errors

```
reg lsalhor exp exp2 age age2 i.diplome i.matrim i.couple immig i.an ENFC90 ENF3, robust
```

- Homogenous model (pooling) and Robust cluster standard errors (by individuals)

```
reg lsalhor exp exp2 age age2 i.diplome i.matrim i.couple immig i.an ENFC90 ENF3, cluster(ident2)
```

Random effects models

- We estimate the following model

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

- There are two random terms a_i and e_{it} with each having their own random distribution.
- Estimation technique is a generalized leas squares (slow!).
- Advantage of this model is that it enables to decompose the residual variance in two components: a between-individual component and a within-individual component (or idiosyncratic).

Implementation: with plm package

```
install.packages("plm")
library(plm)
regre<-plm(lsalhor~exp+exp2+age
  +age2+DDIPL+M+COHAB+immig+ENFC90+ENF3,data=d,
  index=c("ident", "an"), model="random")
summary(regre)
```

With Stata

- Random effects

```
xtset ident2 an
xtreg lsalhor exp exp2 age age2 i.diplome i.matrim i.couple immig i.an ENFC90 ENF3, re
```

Random effects example

```
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
individual    0.09448 0.30737 0.651
theta  :
   Min. 1st Qu.  Median    Mean 3rd Qu.    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 ***
exp          2.0001e-02  1.3143e-03  15.2174 < 2.2e-16 ***
exp2        -2.5480e-04  5.0115e-05  -5.0843 3.722e-07 ***
age          4.0225e-02  5.5856e-03   7.2016 6.160e-13 ***
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 ***
DDIPL7      -7.2696e-01  1.3231e-02 -54.9446 < 2.2e-16 ***
M2          -1.4219e-02  9.0952e-03  -1.5633 0.1179903
M3          -3.4741e-02  3.0758e-02  -1.1295 0.2586942
M4           2.2202e-02  1.2503e-02   1.7757 0.0757952 .
COHAB2      -1.6166e-03  9.3976e-03  -0.1720 0.8634242
immig        1.8494e-03  1.2368e-02   0.1495 0.8811359
ENFC90      -9.0373e-03  3.4866e-03  -2.5920 0.0095494 **
ENF3         3.2673e-02  8.6185e-03   3.7911 0.0001504 ***
```

Comparison of the auto-correlation of errors correction

- The simple OLS pooling *pooling* model supposes that the of variance of residual is the same for all observations and that there's no auto-correlation of residuals

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

- The Huber White correction enables a heterogeneity in residuals variance

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

RCSE and Random effects correction

$$\begin{array}{cc} \sigma_{j1}^2 & \rho_{j12}\sigma_{j1}^2\sigma_{j2}^2 \\ \rho_{j12}\sigma_{j1}^2\sigma_{j2}^2 & \sigma_{j2}^2 \end{array}$$

- RCSE

$$u'u= \begin{array}{c|cccccccccc|}
 & i1 & i2 & i3 & j1 & j2 & j3 & & n1 & n2 & n3 \\
\hline
i1 & \sigma_{i1}^2 & \rho_{i12}\sigma_{i1}^2\sigma_{i2}^2 & \rho_{i13}\sigma_{i1}^2\sigma_{i3}^2 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\
i2 & \rho_{i12}\sigma_{i1}^2\sigma_{i2}^2 & \sigma_{i2}^2 & \rho_{i23}\sigma_{i2}^2\sigma_{i3}^2 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\
i3 & \rho_{i13}\sigma_{i1}^2\sigma_{i3}^2 & \rho_{i23}\sigma_{i2}^2\sigma_{i3}^2 & \sigma_{i3}^2 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\
j1 & 0 & 0 & 0 & \sigma_{j1}^2 & \rho_{j12}\sigma_{j1}^2\sigma_{j2}^2 & \rho_{j13}\sigma_{j1}^2\sigma_{j3}^2 & \ldots & 0 & 0 & 0 \\
j2 & 0 & 0 & 0 & \rho_{j12}\sigma_{j1}^2\sigma_{j2}^2 & \sigma_{j2}^2 & \rho_{j23}\sigma_{j2}^2\sigma_{j3}^2 & \ldots & 0 & 0 & 0 \\
j3 & 0 & 0 & 0 & \rho_{j13}\sigma_{j1}^2\sigma_{j3}^2 & \rho_{j23}\sigma_{j2}^2\sigma_{j3}^2 & \sigma_{j3}^2 & \ldots & 0 & 0 & 0 \\
 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
n1 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & \sigma_{n1}^2 & \rho_{n12}\sigma_{n1}^2\sigma_{n2}^2 & \rho_{n13}\sigma_{n1}^2\sigma_{n3}^2 \\
n2 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & \rho_{n12}\sigma_{n1}^2\sigma_{n2}^2 & \sigma_{n2}^2 & \rho_{n23}\sigma_{n2}^2\sigma_{n3}^2 \\
n3 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & \rho_{n13}\sigma_{n1}^2\sigma_{n3}^2 & \rho_{n23}\sigma_{n2}^2\sigma_{n3}^2 & \sigma_{n3}^2 \\
\hline
\end{array}$$

- Random effects

$$\begin{array}{cc} \sigma_a^2+\sigma_e^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2+\sigma_e^2 \end{array}$$

$$u'u= \begin{array}{c|cccccccccc|}
 & i1 & i2 & i3 & j1 & j2 & j3 & & n1 & n2 & n3 \\
\hline
i1 & \sigma_a^2+\sigma_e^2 & \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\
I2 & \sigma_a^2 & \sigma_a^2+\sigma_e^2 & \sigma_a^2 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\
I3 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma_e^2 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\
j1 & 0 & 0 & 0 & \sigma_a^2+\sigma_e^2 & \sigma_a^2 & \sigma_a^2 & \ldots & 0 & 0 & 0 \\
j2 & 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2+\sigma_e^2 & \sigma_a^2 & \ldots & 0 & 0 & 0 \\
j3 & 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma_e^2 & \ldots & 0 & 0 & 0 \\
 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
n1 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & \sigma_a^2+\sigma_e^2 & \sigma_a^2 & \sigma_a^2 \\
n2 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & \sigma_a^2 & \sigma_a^2+\sigma_e^2 & \sigma_a^2 \\
n3 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma_e^2 \\
\hline
\end{array}$$

Eliminating between-individual variance and focusing on within-individual variance

… and solving the (constant) unobserved heterogeneity problem

The unobserved heterogeneity problem

- Let's go back to random effect model:

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

- If $\mathrm{cov}(x_{kit}, a_i) = 0$, then no problems.
- But if random error a_i can be seen as the product of all time invariant unobservable variables… then it's not impossible that these unobservable variables are correlated with the observable ones used in the model.
- Rather than modeling a_i with random effects models, we can model them with fixed effects.

Solution 1: fixed effects

- 1a) Least square dummy variables LSDV
  - Introduce one dummy variable a_i per individual

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

  - Possible only if the number of individuals is small (<1000).
- 1b) Within models
  - Model (without intercept) where we "demean" each variable (both dependent and independent). We calculate the gap of the variable and its mean for all its observations per individual.

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

- Two methods are equivalent

Fixed effects estimation with R

- With LSDV:

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

- With *within* estimator
  - package plm

```
regfew<-plm(lsalhor~exp+exp2+age+age2+DDIPL
  +M+COHAB+immig+ENFC90+ENF3,
  index=c("ident","an"),model="within",data=d)
summary(regfew)
```

- With lfe package

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

Solution with felm

```
#install.packages("lfe")
library("lfe")
poo<-felm(lsalhor~exp+exp2+age+age2+DDIPL+M
        +COHAB+immig+ENFC90+ENF3|ident|0|0,
        data=d)
```

And with Robust cluster standard errors

```
poo<-felm(lsalhor~exp+exp2+age+age2+DDIPL+M
        +COHAB+immig+ENFC90+ENF3|ident|0|ident,
        data=d)
```

Fixed effects example with R

```
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
```

With Stata

- fixed effects

```
xtset ident2 an
xtreg lsalhor exp exp2 age age2 i.diplome i.matrim i.couple immig i.an ENFC90 ENF3, fe
```

The two fixed effects: group and time

- We can introduce a group fixed effect (ex. individual, country, etc.) and a time fixed effect

==> the common temporal (business) trend captured by the time fixed effects

– All with dummy variables :

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

– *Within* + time dummy variables

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

w2<-plm(lsalhor~exp+…+ENF3+factor(an), index=c("ident","an"), model="within", data=d)

– *Two ways within (formula for balanced panel only)*

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

w2<-plm(lsalhor~exp+…+ENF3, index=c("ident","an"), model="within", effect = "twoways", data=d) # SLOW !!!

Solution 2: First Difference Models

- Explaining evolutions with evolutions

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

- Lag choice
  - [t,t-1]➔ short term variations
  - [t,t-k]➔ long term variations… but less observations
- Difference between fixed effects and first differences
  - Two periods panel: models are equivalent (if the *within* model also has a time fixed effect).
  - Panel with more than two periods: fixed effects model is an average of short and long term evolutions

First difference models

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

Family Status	Model 1 (1–2 Years)	Model 2 (2–4 Years)	Model 3 (3–5 Years)	Model 4 (5–9 Years)
	Model and Number of Years between Observations			
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)

First difference models estimation with softwares

- With R

```
fd<-plm(lsalhor~exp+exp2+age+age2+DDIPL
    +M+COHAB+immig
    +ENFC90+ENF3,
    index=c("ident","an"),model="fd",data=d)
summary(fd)
```

- With stata

```
xtset ident2 an
reg D.(lsalhor exp exp2 age age2 immig ENFC90 ENF3) if  (sexe==2 )
```

  - Warning, it works only with numerical variables and not qualitative variables ==> create dummy variables

First difference

```
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
```

Fixed effects and first difference models. Remarks

- Double effect: reduction of the auto-correlation problem (but not completely suppressed so clustering is welcome) and suppression of the time-invariant unobserved heterogeneity.
- Interpretation : evolutions explain evolutions.
- Disappearance of the intercept in fixed effects models but not in first differences

Fixed effects and first differences models. Remarks (2)

- Disappearance of all time invariant variables ➔ contained in all dummy variables.
  - Ex: migration
  - Sometimes time invariant variables may be estimated because of special cases:
    - Incoherence in responses from one wave of the survey to the other
    - Change in gender
    - Late diploma…
- We can introduce time-invariant variables through interactions (with time or with another key independent variable)
  - Interpretation is still in evolution

```
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
an         6.513e-02  2.346e-02   2.777 0.005497 **
immig:an   1.358e-02  8.548e-03   1.589 0.112169
```

Concentrating on between-individual variance and eliminating within-individual variance

How to study between-individual variance only?

- Fixed effects (or *within*) models enable to study evolutions within-individuals
- Pooled regression and random effects models are two ways of combining within-individual and between-individual variances
- Can we restrict to between-individual variations?
- Yes, two ways
  - Simple single-period regression
  - *Between* Regression

Between regression

- A *between* model is a regression where we regress the mean of the dependent variable per individual on the means of independent of the independent variables per individual.

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{1i} + \ldots + \beta_k \bar{x}_{ki} + \bar{u}_i$$

- This regression informs on between-individual variations.
- It could be the correct one:
  - If the within-individual time variation is neglectable (or random or meaningless)
  - If individuals with within-individual variations are not representative.

With softwares

- R

```
be<-plm(lsalhor~exp+exp2+age+age2+DDIPL
    +M+COHAB+immig
    +ENFC90+ENF3,
    index=c("ident","an"),model="be",data=d)
summary(be)
```

- Stata

```
xtset ident2 an
xtreg lsalhor exp exp2 age age2 i.diplome i.matrim i.couple immig i.an ENFC90
    ENF3, be
```

Between example

```
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 ***
exp          2.3898e-02  1.6630e-03  14.3701 < 2.2e-16 ***
exp2        -2.5506e-04  6.4273e-05  -3.9684 7.283e-05 ***
age          4.2127e-02  6.1914e-03   6.8041 1.071e-11 ***
age2        -4.7735e-04  8.1889e-05  -5.8292 5.731e-09 ***
DDIPL3      -2.2276e-01  1.3685e-02 -16.2771 < 2.2e-16 ***
DDIPL4      -4.1411e-01  1.3801e-02 -30.0071 < 2.2e-16 ***
DDIPL5      -5.8193e-01  1.2809e-02 -45.4316 < 2.2e-16 ***
DDIPL6      -5.7433e-01  1.6899e-02 -33.9855 < 2.2e-16 ***
DDIPL7      -7.0636e-01  1.3721e-02 -51.4807 < 2.2e-16 ***
M2          -1.9630e-02  1.1126e-02  -1.7644   0.07770 .
M3          -2.9286e-02  3.5068e-02  -0.8351   0.40367
M4           3.4091e-02  1.5360e-02   2.2195   0.02647 *
COHAB2      -4.9084e-03  1.1094e-02  -0.4424   0.65817
immig        1.5252e-02  1.2769e-02   1.1945   0.23232
ENFC90      -7.5518e-03  3.9287e-03  -1.9222   0.05461 .
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
```

Links between methods and comparison of their accurateness

Random effects
between fixed effects and *between* models

- There's a mathematical relation between random effect regression, and fixed effects and between regressions:
- $\beta_{Random} = \Delta\, \beta_{Within} + (I\text{-}\Delta)\, \beta_{Between}$
  - Where Δ is a complicated matrix …

Random effects
between fixed effects and *pooling* regression

- A λ parameter measures also the proximity between random effects model, and fixed effects and *pooling* regression.
  - $\lambda=1-[\psi]^{0.5}$ Where $\psi=\sigma_e^2/(\sigma_e^2+T.\ \sigma_a^2)$
  - When ψ converges to 0 (and thus λ to 1) random effects behave as the fixed effects estimator (*within*) based on within-individual variance .
  - When ψ converges to 1 (and thus λ to 0) random effects behaves like the pooling estimator which models total variance, hence both within and between individual variance
- The introduction of individual random effects enables to have an intermediary specification between the model without individual effect and the fixed effects models.
- The hypothesis of a common random distribution of individual effects enable to adopt a model that is neither fully homogenous nor heterogenous.

Random effect as a combination of the fixed effects model and the *pooling* model

- Random effect as

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

- Instead of total "within transformation", we get rid of only a fraction of the within variation

✂➔ GLS Quasi-demeaned estimator

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

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

Example

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

		Random
Enfc90		-0.0090**
Enf3		0.0327***

Panel regressions with simple descriptive statistics

Regression equivalent	Variable	No child <4	At least 1 child <4 an	Difference
Between	Between individuals	3.85	3.89	+4.6%**
Pooling	Between and within individuals	3.87	3.91	+3.8%***
Random	Between and mostly individuals	3.85	3.86	+1.1
Within	Within individuals			-0.5%

Regression equivalent	Variable	Exit Child <4	No change	Entry child <4	Difference
First Difference	Within individuals	0.068	0.051	0.062	-0.3%

Random effects as an intermediary model

- Random effects model is a hybrid model: An intermediary model between fixed effects and *pooling* models
  - If T is important, random effects resemble fixed effects
  - If the unit effect is small (with a small variance), random effects resemble the *pooling* model
  - If the unit effect is strong (big divergence between units), random effects resemble aux fixed effects
  - Random effect models with generalized least squares are better for surveys with big N
    - Property for small N and large T not well known
    - Beck and Katz advise not to use it
      - Or, if we are to use these models, rather with log likelihood estimation method

When should we use random effects?

- When we think of that unit effects are not correlated with the independent variables X
  - Either for theoretical reasons …
  - Or, because you have good control variables
    - Ex: Supra-regional variables for a country-based regression …
- When your interested in the effect to time invariant variables
- When you are interested in differences between units
- When there’s little variations
- When Hausman test tells you results are similar

Fixed effects or random effects?

- How to choose?
- Hausman Test: we look whether fixed effect estimation differ significantly from the random effects one.
- If Hausman test is significant, normally we should prefer fixed effects model.
- If Hausman test is not significant, rather use the random effects model

Hausman test with R

```
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
```

Hausman test with Stata

```
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
```

Panel regression with SAS

- First sort data

```
Proc sort data=a;
By ident an;
Run;
```

- Sas 9.2 and beyond Proc panel

```
Proc panel data=a ranone rantwo btwng pooled fixone fixtwo;
Id ident an;
Model y=x1 x2;
Run;
```

  - Ranone : random individual effects. Rantwo : random individual and temporal effects. Btwng : between. Pooled: Pooled. Fixone :individual fixed effects. Fixtwo : individual and temporal fixed effects.

- Within, with GLM. Avoids the estimation of individual fixed effects!

```
Proc glm data=a;
Absorb ident;
Model y=x1 x2;
Run;
```

Fixed effects beyond panel data

- Fixed effects => etablishments, classes, etc.
- Restrict the space of comparison.
  - Ex. Academic inbreeding
- Fixed effect models have a difference in difference spirit
- Beware to the data restriction/selection due to the field of comparison:

  Observation on which there's no change at all are not used in the analysis

  There might be a selection to estimate things on changers only

Within variations of: the dependent variable The independent variable	Yes	No
Yes	Useful observations	Indirectly useful observation
No	Indirectly useful observation	Unused observations

Limits of fixed effects

- We estimate effects only on individuals which change (on dependent or independent)
  - If y_i and x_i don't change with t, then y_i is captured by the fixed effects a_i.
  - These evolutions, especially with qualitative variables, can be estimated on very specific and singular individuals: selection bias
  - Independent variable changes can also be very specific
    - Extreme case: change in gender.

      Using would be to make the hypothesis that difference in outcome between genders are well estimated by the evolution of outcome for those who change gender
    - Classical: promotion non-manager → manager (*cadre* in France)
  - Fixed effects models lead to considerer that evolutions estimated on people who change are representative of state differences between people who don't change.

Limits of fixed effects

- Yes, fixed effects enable to correct unobserved heterogeneity! But only part of it
  - The time invariant constant one…
- It … transforms a level-regression in an evolution-regression…
  - Where an unobserved factor of evolution correlated to the observed ones used in the regression could still bias the regression
  - Where there could still be simultaneity in dependent and independent variables evolutions. And this could also lead to a bias
  - ←➔ Instrumental variables (if we can find some)
- No correction of time-varying unobserved heterogeneity.
- No correction for reverse causality

Limits of fixed effects

- Essentialist ontology
  - Individual effect: Unchanged, permanent, always the same component
- Evolutions explain evolutions, but the invariant does not account for change (unless contrary specification)

Logistic extension

Extension des panels à des models non lineaires

- Probit (random effects), Logistic (fixed effects, random effects)
  - NB: probit models don't work well with fixed effects
- Package pglm :
  - binomial models (logit and probit), count models (poisson and negbin) and ordered models (logit and probit)
  - Syntax (example) :

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

    - For the moment, the within and between method don't work for logit and probit models with pglm.
- Package survival: conditional logit – equivalent of within for OLS
  - clogit

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

Panel Logit with STATA

- Random effects

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

- Fixed effects

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

- Hausman test

```
hausman fixed random
```

Panel Logit with SAS

- Fixed effects

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

- Effets aleatoires
  - Proc glimmix

Safi, Rathelot
Is there a white flight in 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

- Apparently yes.
  - Natives move all the more that migrants are in important number
- But do they move because migrants or because of unobserved characteristics neighborhood correlate to migrants’ presence?

Introduction of a municipality fixed effect

- No more white flight
- A tendency to migrant isolation

	(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	.018	.022				
Immigrants						
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	

But in the appendix

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	142,652		611,486	
Pseudo R-Sq	.14			

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

Explanations

- "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: we follow movers on several municipalities

References

Waldfogel Jane, 1997, "The effect of Children on woman's wage", *American Sociological Review.*

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