Econometrics 1
Lecture 9: Panel data
ENSAE 2014/2015

Michael Visser (CREST-ENSAE)

Panel data

In today's lecture we continue with the study of panel data.

Recall that panel data are data where we have multiple observations for each unit (individual, firm, country, etc.) over time.

Panel data have several advantages (over cross-section data):

- More observations (which improves the precision of estimators)
- Additional source of variation (over individuals *and* over time)
- Panel data methods allow to obtain estimators that are robust to certain types of omitted variable bias
- Learn about dynamics

In the following we assume

- Random sampling: the sample of units is randomly drawn from the population (in the cross-section)
- Balanced panel: $T_i = T$ (number of observations per unit i is T for each i)
- Large N (number of units), small T (number of observations per unit)

Fixed effects and random effects methods

The panel data model we study throughout the lecture has the following form

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + \underbrace{a_i + e_{it}}_{u_{it}} = \beta_0 + x_{it}\beta + \underbrace{a_i + e_{it}}_{u_{it}} \quad (1)$$

for $i = 1, \ldots, N$ individuals over $t = 1, \ldots, T$ time periods. Here x_{it} is a $1 \times k$ vector and β a $k \times 1$ vector. We also assume throughout that the following condition holds:

$$E[e_{it}|a_i, x_{i1}, \ldots, x_{iT}] = 0, \qquad t = 1, \ldots, T \quad (2)$$

When this condition holds we say that the explanatory variables $x_{i1}, \ldots, x_{iT}$ are *strictly exogenous conditional on the unobserved effect* a_i.

Two types of estimation methods will be studied, depending on the assumption made on $E[a_i|x_{it}]$:

Fixed effects and random effects methods (cntd)

- Fixed Effects (FE) methods. In this setup the unit-specific effect a_i may be correlated with x_{it}. In this case we have

$$E[y_{it}|x_{it}] = \beta_0 + x_{it}\beta + \underbrace{E[a_i|x_{it}]}_{\text{unknown}}$$

- Random Effects (RE) methods. In this setup a_i must be uncorrelated with x_{it}: $E[a_i|x_{it}] = E[a_i] = 0$ (normalizing the unconditional mean of a_i to 0 is without loss of generality as long as there is a constant in (1)). We have in this case

$$E[y_{it}|x_{it}] = \beta_0 + x_{it}\beta + \underbrace{E[a_i|x_{it}]}_{=0} = \beta_0 + x_{it}\beta$$

If we can assume that a_i and x_{it} are uncorrelated for each t then RE methods are the most appropriate. If there is correlation then the FE methods should be used.

Fixed Effects

Consider model (1) with just one explanatory variable:

$$y_{it} = \beta_0 + x_{it1}\beta_1 + a_i + e_{it}$$

Pooling all observations and regressing the model by OLS leads to a biased estimator of β_1 if $E[a_i|x_{it1}] \neq 0$.

To illustrate this bias, suppose there are two groups of individuals. In the first group, individuals always have "small' values for their explanatory variables, and in the second group the regressors are always "large".

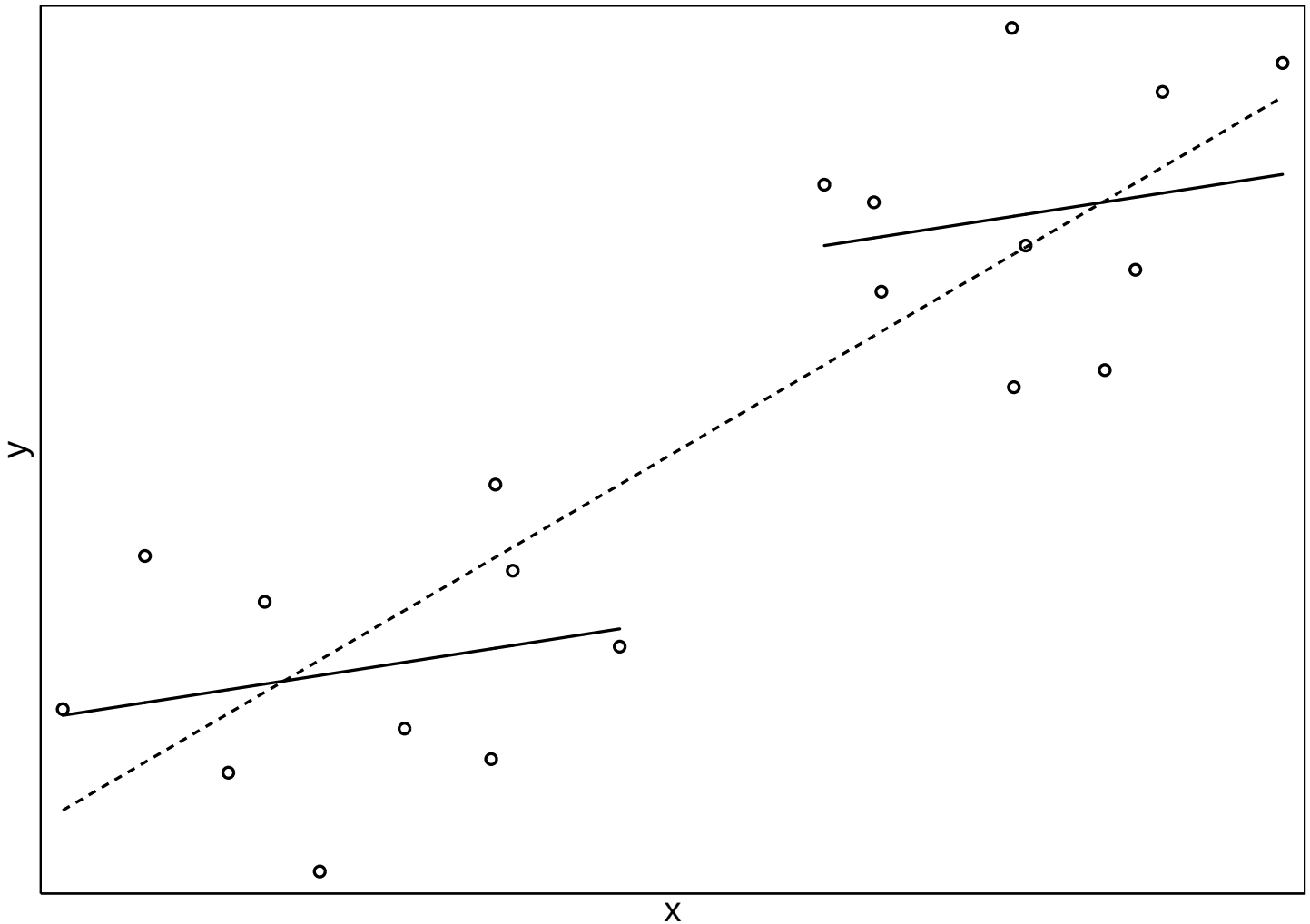
For all individuals i in the first group the expected value of a_i is "small", and for all i in the second group this expectation is "large":

$$E[a_i|x_{it1} \textit{ is large}] > E[a_i|x_{it1} \textit{ is small}]$$

Suppose that $E[a_i|x_{it1} \textit{ is small}] > 0$, then the OLS estimator of the slope paramater β_1 is biased upwards (see next graph).

Fixed Effects

Illustration of bias



Fixed Effects

First difference estimator

As we saw in last week's lecture (for $T = 2$) differencing model (1) removes the fixed effects:

$$\begin{aligned} y_{it} - y_{i,t-1} &= (x_{it} - x_{i,t-1})\beta + (e_{it} - e_{i,t-1}) \\ \Delta y_{it} &= \Delta x_{it}\beta + \Delta e_{it},\ i = 1, ..., N,\ t = 2, ..., T \end{aligned}$$

In differencing we lose the first observation for each i and we cannot identify the constant β_0.

The *first difference* estimator, $\hat{\beta}_{FD}$, is the pooled OLS estimator from the regression of Δy_{it} on Δx_{it}, $i = 1, ..., N,\ t = 2, ..., T$.

This estimator is consistent (under any form of the correlation between a_i and x_{it}) if $E[\Delta e_{it}|\Delta x_{it}] = 0$, which holds under the strict exogeneity assumption.

To determine the asymptotic variance of the estimator of the estimator, we assume in addition that

$$E[\Delta e_i(\Delta e_i)'|a_i, x_{i1}, \ldots, x_{iT}] = \sigma^2_{\Delta e} I_{T-1}$$

Fixed Effects

First difference estimator

where Δe_i is the $(T-1) \times 1$ vector containing $e_{it} - e_{it-1}$, $t = 2, ..., T$, and I_{T-1} is the identity matrix of dimension $T-1$.

Under this assumption the differences $\Delta e_{it} = e_{it} - e_{it-1}$ are serially uncorrelated, and have constant variance $\sigma^2_{\triangle e}$.

Using the same proofs as in previous lectures (3, 4 and 5), the estimator $\hat{\beta}_{FD}$ is asymptotically normally distributed, and its asymptotic variance can be estimated by (in matrix notation):

$$\widehat{aVar}(\hat{\beta}_{FD}) = \hat{\sigma}^2_{\triangle e}(\triangle X' \triangle X)^{-1}$$

where $\triangle X$ is a $N(T-1) \times k$ matrix that stacks all the $N(T-1)$ vectors $\triangle x_{it}$ one under the other, and $\hat{\sigma}^2_{\triangle e}$ is a consistent estimator of $\sigma^2_{\triangle e}$:

$$\hat{\sigma}^2_{\Delta e} = [N(T-1) - k]^{-1} \sum_{i=1}^{N} \sum_{t=2}^{T} (\triangle y_{it} - \triangle x_{it} \hat{\beta}_{FD})^2$$

Fixed Effects

Within Estimator

First differencing is one method to eliminate the unit-specific effects a_i.

Another method is the *fixed effects* transformation. It works as follows.

First take averages over time in model (1)

$$\bar{y}_i = \beta_0 + \bar{x}_i\beta + a_i + \bar{e}_i$$

where

$$\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}; \quad \bar{x}_i = T^{-1}\sum_{t=1}^{T} x_{it}; \quad \bar{e}_i = T^{-1}\sum_{t=1}^{T} e_{it}.$$

Next take deviations from time means to obtain the centered model

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + (e_{it} - \bar{e}_i)$$

or more succinctly

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + \ddot{e}_{it} \tag{3}$$

Fixed Effects

Within Estimator

The effects a_i have dropped out from the centered model (and the intercept β_0 as well).

The *within* estimator, $\hat{\beta}_{Within}$, is obtained by estimating (3) with pooled OLS. That is, we regress $\ddot{y}_{it}$ on $\ddot{x}_{it}$ using all observations $i = 1, ..., N$, $t = 1, ..., T$.

The within estimator is also called the *fixed effects* estimator.

As with the first difference estimator, the vector of parameters β is only identified if the regressors x_{it} vary over time.

For $\hat{\beta}_{Within}$ to be a consistent estimator of β we need the error to be orthogonal to the regressors

$$E[\ddot{x}_{it}\ddot{e}_{it}] = E[(x_{it} - \bar{x}_i)(e_{it} - \bar{e}_i)] = 0$$

which follows from the strict exogeneity assumption (2).

Assumption (2) rules out models with feedback (where a regressor x_{itj} is correlated with e_{it-1}), lagged dependent variables, and other types of "endogenous" regressors.

To see why lagged dependent variables cannot be included, consider the model

$$y_{it} = x_{it}\beta + a_i + e_{it}$$

where $x_{it} = y_{i,t-1}$ (so x_{it} is a scalar). Assume that the following holds

$$E[e_{it}|a_i, x_{i1}, \ldots, x_{it}] = 0$$

then we see that

$$E[x_{i,t+1}e_{it}] = E[y_{it}e_{it}] = E[x_{it}e_{it}]\beta + E[a_i e_{it}] + E[e_{it}^2] = E[e_{it}^2] > 0$$

and this implies $E[e_{it}|x_{i,t+1}] \neq 0$, so the strict exogeneity condition (2) is violated.

Fixed Effects

Within Estimator

Under $E[e_i e_i' | a_i, x_{i1}, \ldots, x_{iT}] = \sigma_e^2 I_T$ (e_i is now a $T \times 1$ vector containing e_{it}, $t = 1, ..., T$), the estimator $\hat{\beta}_{Within}$ is asymptotically normally distributed, and its asymptotic variance can be estimated by:

$$\widehat{aVar}(\hat{\beta}_{Within}) = \hat{\sigma}_e^2 (\ddot{X}' \ddot{X})^{-1} \tag{4}$$

where $\ddot{X}$ is a $NT \times k$ matrix that stacks all the NT vectors $\ddot{x}_{it}$ one under the other, and $\hat{\sigma}_e^2$ is a consistent estimator of σ_e^2 :

$$\hat{\sigma}_e^2 = [N(T-1) - k]^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} (\ddot{y}_{it} - \ddot{x}_{it} \hat{\beta}_{Within})^2 \tag{5}$$

The estimator $\hat{\beta}_{Within}$ is simply the OLS estimator of β in the transformed model (3).

Fixed Effects

Within Estimator

The usual OLS procedures would produce an estimated asymptotic variance of the form (4), but it would wrongly assume that there are $NT - k$ degrees of freedom.

The mistake that one makes by taking the naive df $NT - k$ (instead of $N(T-1) - k$) can be important if T is relatively small.

The true df is $N(T-1) - k$ because of the time-demeaning operation that led to model (3). For each i the sum of demeaned errors $\sum_{t=1}^{T} \ddot{e}_{it} = 0$, so for each i we lose one df.

Fixed Effects

Within estimator

It is straightforward to estimate the variance of the fixed effects a_i.

The variance of u_{it} can be estimated by

$$\hat{\sigma}_u^2 = \frac{1}{NT - k} \sum_i \sum_t (y_{it} - x_{it}\hat{\beta}_{Within})^2$$

Since $\sigma_u^2 = \sigma_a^2 + \sigma_e^2$ ($Var(u_{it}) = Var(a_i) + Var(e_{it})$ since $Cov(a_i, e_{it}) = 0$ under (2)), we can estimate the variance of a_i by

$$\hat{\sigma}_a^2 = \hat{\sigma}_u^2 - \hat{\sigma}_e^2$$

where $\hat{\sigma}_e^2$ is defined in (5). We can also calculate the predicted fixed effects:

$$\hat{a}_i = \bar{y}_i - \bar{x}_i\hat{\beta}_{Within}$$

Many econometric software packages (such as Stata) also report the following estimate of the intercept β_0 :

$$\hat{\beta}_0 = \frac{1}{N} \sum_{i=1}^{N} \hat{a}_i$$

Fixed Effects

Dummy variable regression

We can also obtain the within estimator via estimation by OLS of the following model without constant

$$y_{it} = x_{it}\beta + \sum_{i=1}^{N} \alpha_i d_{it} + e_{it}$$

where d_{it} are dummy variables that equal 1 for individual i (zero otherwise). It can be shown that the OLS estimators of β and α_i, $i = 1, ..., N$, are in fact equivalent to the within estimators $\hat{\beta}_{Within}$ and $\hat{a}_i$, $i = 1, ..., N$.

The DV regression is not convenient for large datasets, i.e., datsets with many units (there are $k + N$ parameters to be estimated and this number is large for large N).

WE or FD?

When should one use the within estimator (WE) and when the first difference estimator (FD)?

- When $T = 2$, it doesn't matter since they are identical
- $T > 2$, FD$\neq$WE
  - If e_{it} are serially uncorrelated: FD less efficient than WE
  - If e_{it} are serially correlated: FD generally more efficient than WE
  - FD is more sensitive to violations of strict exogeneity
  - If WE and FD differ substantially then this suggests that the strict exogeneity does not hold

Random Effects

Consider again the model

$$y_{it} = \beta_0 + x_{it}\beta + \underbrace{a_i + e_{it}}_{u_{it}}$$

We still assume that the strict exogeneity assumption holds:

$$E[e_{it}|a_i, x_{i1}, \ldots, x_{iT}] = 0, \qquad t = 1, \ldots, T$$

But now we assume in addition that

$$E[a_i|x_{it}] = 0, \qquad t = 1, \ldots, T$$

Estimation of the above regression model by pooled OLS gives unbiased estimators of β_0 and β since the two assumptions on the error terms imply $E[u_{it}|x_{it}] = 0$.

Random Effects

If we assume in addition that

$$E[e_i e_i' | a_i, x_{i1}, \ldots, x_{iT}] = \sigma_e^2 I_T, \qquad E[a_i^2 | x_{i1}, \ldots, x_{iT}] = \sigma_a^2$$

then

$$\Omega \equiv E[u_i u_i'] = \begin{pmatrix} \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \cdots & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & & \vdots \\ \vdots & & \ddots & \sigma_a^2 \\ \sigma_a^2 & \cdots & \cdots & \sigma_a^2 + \sigma_e^2 \end{pmatrix}$$

Hence the error terms u_{it} are serially correlated over time. Indeed, for $s \neq t$ we have

$$corr(u_{is}, u_{it}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} > 0$$

Since the usual pooled OLS standard errors ignore this positive correlation, they will be incorrect, as will the usual test statistics.

Using FGLS we can get rid of the serial correlation problem here.

Random Effects

Write the T observations of unit i in vector form:

$$y_i = x_i^* \beta^* + u_i$$

where y_i is the $T \times 1$ vector that stacks all y_{it} one under the other, x_i^* is the $T \times (k+1)$ matrix that stacks the $(k+1)$-dimensional vectors x_{it}^* one under the other (with $x_{it}^* = (1\, x_{it1} \,..x_{itk})$), $\beta^* = (\beta_0\, \beta_1 \,... \,\beta_k)$' (a $(k+1) \times 1$ vector), and u_i is the $T \times 1$ vector that stacks all u_{it} one under the other.

Now consider the transformed model

$$\Omega^{-1/2} y_i = \Omega^{-1/2} x_i^* \beta^* + \Omega^{-1/2} u_i \tag{6}$$

where $\Omega^{-1/2}$ is the matrix such that $\Omega^{-1/2}\Omega^{-1/2} = \Omega^{-1}$, the latter matrix being the inverse matrix of Ω. We have

$$E[\Omega^{-1/2} u_i u_i' \Omega^{-1/2}] = \Omega^{-1/2}\Omega\Omega^{-1/2} = I_T$$

Random Effects

Hence the error terms in the transformed model are no longer serially correlated.

Letting $\hat{\Omega}$ be a consistent estimator of Ω, the *random effects* estimator is the FGLS estimator of β^*. It is obtained via estimation by pooled OLS of model (6) after replacing Ω by $\hat{\Omega}$:

$$\hat{\beta}^*_{RE} = (\sum_{i=1}^{N} x_i^{*'} \hat{\Omega}^{-1} x_i^*)^{-1} (\sum_{i=1}^{N} x_i^{*'} \hat{\Omega}^{-1} y_i)$$

To estimate Ω use OLS fitted residuals $\hat{u}_{it} = y_{it} - x^*_{it}\hat{\beta}^*_{OLS}$:

$$\widehat{E[u_{it}^2]} = \widehat{\sigma_a^2 + \sigma_e^2} = \frac{1}{NT - k - 1} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{it}^2$$

and for $t \neq s$

$$\widehat{E[u_{it} u_{is}]} = \widehat{\sigma_a^2} = \frac{1}{NT(T-1)/2 - k - 1} \sum_{i=1}^{N} \sum_{t,s:t\neq s} \hat{u}_{it} \hat{u}_{is}$$

Random Effects

The estimator $\hat{\beta}^*_{RE}$ is asymptotically normally distributed, and its asymptotic variance can be estimated by:

$$\widehat{aVar}(\hat{\beta}^*_{RE}) = (\sum_{i=1}^{N} x_i^{*\prime} \hat{\Omega}^{-1} x_i^*)^{-1} \tag{7}$$

Example - Labor supply

Suppose we are interested in estimating a labor supply equation

$$\ln(hours_{it}) = \beta_0 + \beta \ln(wage_{it}) + a_i + e_{it}$$

```
. sum

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          id |       5320       266.5    153.5893         1        532
        year |       5320      1983.5    2.872551      1979       1988
        lnhr |       5320     7.65743    .2855914      2.77       8.56
        lnwg |       5320    2.609436    .4258924      -.26       4.69
        kids |       5320    1.555827    1.195924         0          6
        ageh |       5320    38.91823    8.450351        22         60
       agesq |       5320    1586.024    689.7759       484       3600
       disab |       5320    .0609023    .2391734         0          1
```

Example - Labor supply

```
. list, sep(10)

     +-------------------------------------------------------------+
     |  id   year   lnhr   lnwg   kids   ageh   agesq   disab |
     |-------------------------------------------------------------|
  1. |   1   1979   7.58   1.91      2     27     729       0 |
  2. |   1   1980   7.75   1.89      2     28     784       0 |
  3. |   1   1981   7.65   1.91      2     29     841       0 |
  4. |   1   1982   7.47   1.89      2     30     900       0 |
  5. |   1   1983    7.5   1.94      2     31     961       0 |
  6. |   1   1984    7.5   1.93      2     32    1024       0 |
  7. |   1   1985   7.56   2.12      2     33    1089       0 |
  8. |   1   1986   7.76   1.94      2     34    1156       0 |
  9. |   1   1987   7.86   1.99      2     35    1225       0 |
 10. |   1   1988   7.82   1.98      2     36    1296       0 |
     |-------------------------------------------------------------|
 11. |   2   1979    7.2   2.54      4     35    1225       0 |
 12. |   2   1980   6.95   2.52      3     37    1369       1 |
 13. |   2   1981   7.24   2.59      3     37    1369       1 |
 14. |   2   1982   7.46   2.51      3     38    1444       1 |
 15. |   2   1983   6.81   2.77      3     39    1521       0 |
 16. |   2   1984   5.44   1.43      2     40    1600       0 |
 17. |   2   1985   5.08   1.72      1     42    1764       1 |
 18. |   2   1986   5.85   1.86      1     42    1764       0 |
 19. |   2   1987   7.69   1.83      1     43    1849       0 |
 20. |   2   1988   7.63   1.79      0     44    1936       0 |
     |-------------------------------------------------------------|
 21. |   3   1979   7.43   2.39      0     26     676       1 |
 22. |   3   1980   7.25   2.38      0     27     729       0 |
 23. |   3   1981   7.62   2.58      0     28     784       0 |

[etc]
```

Example - Labor supply

```
. xtreg lnhr lnwg, re i(id)

Random-effects GLS regression                   Number of obs      =      5320
Group variable: id                              Number of groups   =       532

R-sq:  within  = 0.0162                         Obs per group: min =        10
       between = 0.0213                                        avg =      10.0
       overall = 0.0152                                        max =        10

Random effects u_i ~ Gaussian                   Wald chi2(1)       =     76.64
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        lnhr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnwg |   .1193322   .0136312     8.75   0.000     .0926155     .146049
       _cons |   7.346041   .0363925   201.86   0.000     7.274713    7.417368
-------------+----------------------------------------------------------------
     sigma_u |  .16124733
     sigma_e |  .23278339
         rho |  .32424354   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. estimates store RE
```

Relationship RE, FE, and OLS estimators

It can be shown that the transformed model (6) implies the following quasi-differenced model (after replacing Ω by $\hat{\Omega}$)

$$y_{it} - \hat{\lambda}\bar{y}_i = \beta_0(1 - \hat{\lambda}) + (x_{it} - \hat{\lambda}\bar{x}_i)\beta + (u_{it} - \hat{\lambda}\bar{u}_i) \qquad (8)$$

where

$$\hat{\lambda} = 1 - \frac{\hat{\sigma}_e}{\sqrt{\hat{\sigma}_e^2 + T\hat{\sigma}_a^2}}$$

and $\bar{y}_i = T^{-1}\sum_{i=1}^{N} y_{it}$, $\bar{x}_i = T^{-1}\sum_{i=1}^{N} x_{it}$, and $\bar{u}_i = T^{-1}\sum_{i=1}^{N} u_{it}$.

So the random effects estimator is the OLS estimator of β_0 and β in model (8).

The pooled OLS estimator and the within estimator correspond to the two following cases:

- $\hat{\lambda} = 0$: Pooled OLS
- $\hat{\lambda} = 1$: Within (FE)

Testing

Hausman test

To test $\mathcal{H}_0 : E[a_i|x_{it}] = 0$ against $\mathcal{H}_1 : E[a_i|x_{it}] \neq 0$, Hausman (1978) has devised a test based on the comparison of the within estimator and the random effects estimator. The idea is that

- Under the null hypothesis, both RE and FE estimators are consistent (but the RE estimator is more efficient)
- Under the alternative hypothesis, only the FE estimator is consistent.

This implies that under the null plim $\hat{\beta}_{FE} =$ plim $\hat{\beta}_{RE}$, but under the alternative plim $\hat{\beta}_{FE} \neq$ plim $\hat{\beta}_{RE}$. A substantial difference between the two estimates should therefore be interpreted as evidence against the null. The test statistic is

$$H = (\hat{\beta}_{Within} - \hat{\beta}_{RE})' \left(\widehat{aVar}(\hat{\beta}_{Within}) - \widehat{aVar}(\hat{\beta}_{RE}) \right)^{-1} (\hat{\beta}_{Within} - \hat{\beta}_{RE})$$

Under the null this statistic follows a χ^2_k. The null should be rejected if the outcome of the statistic is larger than the $\alpha\%$ critical value of a χ^2_k.

Example - Labor supply

Hausman test

```
. xtreg lnhr lnwg, fe i(id)
. estimates store FE
. hausman FE RE

                ---- Coefficients ----
             |      (b)           (B)             (b-B)     sqrt(diag(V_b-V_B))
             |       FE            RE         Difference          S.E.
-------------+------------------------------------------------------------------
        lnwg |    .1676755     .1193322        .0483432        .0130486
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       13.73
                Prob>chi2 =      0.0002
```

I understand/can apply...

- The structure of panel data
- Fixed effects methods
  - 1st-difference and within estimator
  - relation to OLS with dummies
  - strict exogeneity
- Random effects methods
  - standard covariance structure
  - FGLS estimator
- Hausman test