# FE and RE Problems

José Fajardo
FGV/EBAPE

## Is There Enough Within Variation?

- FE identifies the causal effect by using within variation only
  - Therefore always ask two questions before doing FE-analysis
  - Is there enough within variation?
    - Without observing enough changes in X, one cannot apply FE
  - What is the treatment population?
    - FE identifies "average treatment effects on the treated" (ATT)
    - ATTs can only be generalized to those, who potentially can experience the change in X

## Endogeneity

- FE-estimator has to assume strict exogeneity

  $$Cov(x_{it}, \varepsilon_{is}) = 0 \quad \text{for all t and s}$$

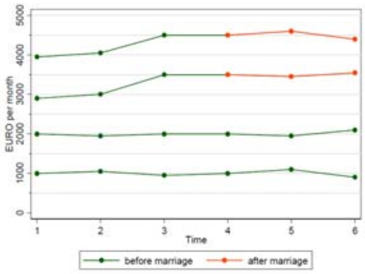
  - Endogeneity: $Cov(x_{it}, \varepsilon_{is}) \neq 0$
  - FE-estimates are biased under endogeneity
    - However, the bias may be very small for large T
- Sources of endogeneity (Wooldridge, 2010: 321)
  - Time-varying omitted variables (captured by $\varepsilon_{it}$)
    - Unobservables that affect both X and Y
  - Y affects also X (simultaneity)
    - There is an exogeneous "shock" on Y, and the change in Y affects X (see example below)
  - Errors in reporting X (measurement errors)
    - Measurement errors generally produce an "attenuation bias". With more X-variables the direction of the bias is unknown, however.

→ **The event (change in X) is no longer exogeneous**

## FE Estimators under Endogeneity

- Instrumental variables (IV) estimation (`xtivreg`)
  - FD-IV and FE-IV are available.
  - If "sequential exogeneity" is maintainable, then FD-IV can use the lagged regressors as a valid instrument
- Structural equation modeling (LISREL)
  - LISREL models supposedly can take regard of simultaneity and measurement errors
- Problems of these methods
  - They rest on untestable assumptions
    - Exogeneity of the instruments
  - They are not robust

## Marriage Example: Simultaneity




Modified Data:

The wage jump at t=3 triggers marriage, i.e., causality runs the other way.

There is no additional causal effect of marriage!

There is still selectivity.

## Marriage Example: Simultaneity

```
. xtreg wage2 marr, fe

Fixed-effects (within) regression               Number of obs      =        24
Group variable: id                              Number of groups   =         4

                                                F(1,19)            =     15.72
corr(u_i, Xb)  = 0.5227                         Prob > F           =    0.0008

------------------------------------------------------------------------------
       wage2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        marr |        350   88.27456     3.96   0.001     165.2392    534.7608
       _cons |       2575   38.22401    67.37   0.000     2494.996    2655.004
-------------+----------------------------------------------------------------
```

Data: Wage Premium.dta
Do-File: Wage Premium.do

- Now the FE-estimator is biased
- The IV-estimator is also biased (+238)
  - `xtivreg` with lagged marriage-dummy as instrument (xtivreg wage2 (marr=L.marr), fe)
- Some Ideas: the FD-estimator would provide the correct answer!
  - Solution I: getting the observation window right (hand-made)
  - Solution II: PS-matching on panels (life-courses)

Mariagepremium3.do

## Cornwell and Rupert Data

**Cornwell and Rupert Returns to Schooling Data, 595 Individuals, 7 Years**
**Variables in the file are**

| | |
|---|---|
| EXP | = work experience |
| WKS | = weeks worked |
| OCC | = occupation, 1 if blue collar, |
| IND | = 1 if manufacturing industry |
| SOUTH | = 1 if resides in south |
| SMSA | = 1 if resides in a city (SMSA) |
| MS | = 1 if married |
| FEM | = 1 if female |
| UNION | = 1 if wage set by union contract |
| ED | = years of education |
| LWAGE | = log of wage = dependent variable in regressions |

.

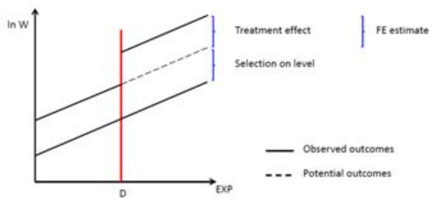
## Exercicio3.do

```
* IV pooled constant effects model
ivregress 2sls lwage exp exp2 (wks = occ ind south smsa ms union)
estimates store IV
estat endogenous
estat overid
estat firststage
* IV fixed effects model
xtivreg lwage exp exp2 (wks = occ ind south smsa ms union), fe
estimates store FE_IV
xtreg lwage exp exp2 wks, fe
estimates store FE_OLS
hausman FE_IV FE_OLS
estimates table IV FE_IV FE_OLS, b se t
```

## The Parallel Trend Assumption

Another way to express the strict exogeneity assumption from a counterfactual perspective

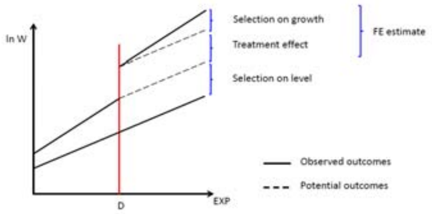
- The (potential) outcome trends in treatment and control group must be parallel
- Then the FE-estimate captures the treatment effect



# The Parallel Trend Assumption

Parallel trend assumption violated

- Those on a steeper trajectory are selected in the treatment
- The FE-estimate will be biased upwards
  - It includes a "steeper growth part"




# What to Do?

- There are FE-models that work with a weaker exogeneity assumption
  - FE-IS (see below)
- Be critical!
  - If you have doubts on the validity of the strict exogeneity assumption, be careful when interpreting FE results
  - Use IV-methods only, when you have good arguments for the validity of the instruments used
    - **Keep it simple**: generally, complex statistical models (LISREL) will not make your inferences better
  - If you find no good IV, be conservative and conclude that even panel data do not help in identifying the causal effect
  - Invest in "**shoe leather**"
    - Try to collect better data that include the time-varying unobservables
    - Try to collect better data by (natural) experiments

# FE with Individual Slopes

- Solution: Fixed-effects model with individual slopes
  - FE-IS: Polachek/Kim 1994; Wooldridge 2010
- FE-IS extends within-transformation of conventional FE
  - Allows for individual slopes in addition to individual constants
  - Idea: Subtract not just mean wage (individual constant), but individual wage profile (individual constant and slope)
  - **xtfeis.ado**: available from the authors
- FE:

$$\ln w_{it} = \alpha_1 exp_{it} + \alpha_2 exp_{it}^2 + \beta_1 m_{it} + \ldots + \alpha_i + \varepsilon_{it}$$

- FE-IS:

$$\ln w_{it} = \alpha_{1i} exp_{it} + \alpha_{2i} exp_{it}^2 + \beta_1 m_{it} + \ldots + \alpha_i + \varepsilon_{it}$$

Exercicio3.do

## CR1988: RE-IV

```
* IV random effects model
xtivreg lwage exp exp2 (wks ed = occ ind south smsa
ms union blk fem), re
estimates store RE_IV
* Random effects model
xtreg lwage exp exp2 wks ed, re
estimates store RE
estimates table IV RE_IV RE, b se t
```

Exercicio4.do

## Hausman-Taylor: xthtaylor

```
xthtaylor lwage occ south smsa ind exp exp2
  wks ms union fem blk ed,
endog(exp exp2 wks ms union ed)
estimates store TH
estimates table IV RE_IV RE TH, b se t
```

Exercicio4.do

## Measurement Error

Standard regression results: General effects model

$y_{it} = x^*_{it}\beta + c_i + \varepsilon_{it}$

$x_{it} = x^*_{it} + h_{it}$

$x_{it} =$ measured variable, including measurement error.

$b = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = (\mathbf{x}'\mathbf{x}/\Sigma_{i=1}^{N} T_i)^{-1} \left\{(\mathbf{x}^* + \mathbf{h})'[\mathbf{x}^*\beta + \mathbf{c} + \boldsymbol{\varepsilon}]/\Sigma_{i=1}^{N} T\right\}$

$$\text{plim } b = \beta\left(\frac{Var[x^*_{it}]}{Var[x^*_{it}] + Var[h_{it}]}\right) + \frac{Cov[x^*_{it}, c_i]}{Var[x^*_{it}] + Var[h_{it}]}$$

biased twice, possibly in opposite directions.

(Griliches and Hausman (1986).)

## Application: A Twins Study

"Estimates of the Economic Returns to Schooling from a New Sample of Twins," Ashenfelter, O. and A. Kreuger, Amer. Econ. Review, 12/1994.

(1) Annual twins study in Twinsburg, Ohio.

(2) (log) wage equations, $y_{i,j} = \log wage$ twin j=1,2 in family i.

(3) Measured data:

(a) Self reported education, Sibling reported education, Twins report same education, other twin related variables

(b) Age, Race, Sex, Self employed, Union member, Married, of mother at birth

(4) $S_j^k =$ reported schooling by of twin j by twin k. Measurement error.

$S_j^k = S_j + v_j^k$. Reliability ratio = $Var[S_j]/(Var[S_j]+Var[v_j^k])]$

## Example: Marriage Premium

* Effect of marr is 500

* Measurement error in the X-variable (marr)

* Pooled OLS

regress wage marr1

* Fixed-Effects Regression (within estimator)

xtreg   wage marr1, fe

## General Conclusions About Measurement Error

- In the presence of individual effects, inconsistency is in unknown directions
- With panel data, different transformations of the data (first differences, group mean deviations) estimate different functions of the parameters – possible method of moments estimators
- Model may be estimable by minimum distance or GMM
- With panel data, lagged values may provide suitable instruments for IV estimation.
- Various applications listed in Baltagi (pp. 187-190).

# Panel Robust Statistical Inference

# Problems

- Heteroscedasticity
- Autocorrelation
- Covariance Structures
- Spatial Autocorrelation

# Heteroscedasticity

- Naturally expected in microeconomic data, less so in macroeconomic
- Model Platforms
  - Fixed Effects $y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$, $E[\varepsilon_{it}^2 \mid \mathbf{X}_i] = \sigma_{\varepsilon,it}^2$
  - Random Effects $y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it}$, $E[\varepsilon_{it}^2 \mid \mathbf{X}_i] = \sigma_{\varepsilon,it}^2$

$$E[u_i^2 \mid \mathbf{X}_i] = \sigma_{u,i}^2$$

- Estimation
  - OLS with (or without) robust covariance matrices
  - GLS and FGLS
  - Maximum Likelihood

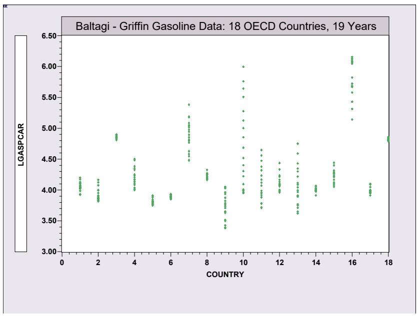
# Baltagi and Griffin's Gasoline Data

**World Gasoline Demand Data, 18 OECD Countries, 19 years**
**Variables in the file are**

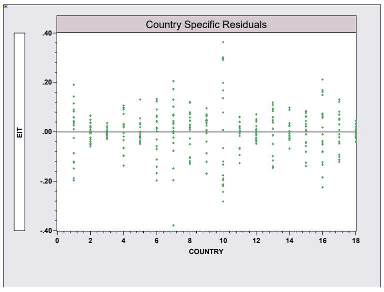
COUNTRY = name of country
YEAR = year, 1960-1978
LGASPCAR = log of consumption per car
LINCOMEP = log of per capita income
LRPMG = log of real price of gasoline
LCARPCAP = log of per capita number of cars

The article on which the analysis is based is Baltagi, B. and Griffin, J., "Gasoline Demand in the OECD: An Application of Pooling and Testing Procedures," European Economic Review, 22, 1983, pp. 117-137.

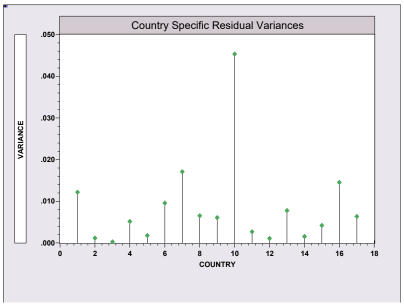
# Heteroscedastic Gasoline Data




# LSDV Residuals



## Evidence of Country Specific Heteroscedasticity




## Heteroscedasticity in the FE Model

- Ordinary Least Squares
  - Within groups estimation as usual.
  - Standard treatment – this is just a (large) linear regression model.
  - White estimator

$$\mathbf{b} = \left[\Sigma_{i=1}^{N}\mathbf{X}_i'\mathbf{M}_D^i\mathbf{X}_i\right]^{-1}\left[\Sigma_{i=1}^{N}\mathbf{X}_i'\mathbf{M}_D^i\mathbf{y}_i\right]$$

$$\text{Var}[\mathbf{b}\,|\,\mathbf{X}] = \left[\Sigma_{i=1}^{N}\mathbf{X}_i'\mathbf{M}_D^i\mathbf{X}_i\right]^{-1}\left[\Sigma_{i=1}^{N}\Sigma_{\mathbf{t=1}}^{T_i}\sigma_{\varepsilon,it}^2(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\right]\left[\Sigma_{i=1}^{N}\mathbf{X}_i'\mathbf{M}_D^i\mathbf{X}_i\right]^{-1}$$

White Robust Covariance Matrix Estimator

$$\text{Est.Var}[\mathbf{b}\,|\,\mathbf{X}] = \left[\Sigma_{i=1}^{N}\mathbf{X}_i'\mathbf{M}_D^i\mathbf{X}_i\right]^{-1}\left[\Sigma_{i=1}^{N}\Sigma_{\mathbf{t=1}}^{T_i}e_{it}^2(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\right]\left[\Sigma_{i=1}^{N}\mathbf{X}_i'\mathbf{M}_D^i\mathbf{X}_i\right]^{-1}$$

$$\underbrace{\begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,T} \\ y_{2,1} \\ \vdots \\ y_{2,T} \\ \vdots \\ \vdots \\ y_{n,1} \\ \vdots \\ y_{n,T} \end{pmatrix}}_{=y} = \underbrace{\begin{pmatrix} 1 & & & \\ \vdots & & & \\ 1 & & & \\ & 1 & & \\ & \vdots & & \\ & 1 & & \\ & & \ddots & \\ & & & \ddots \\ & & & 1 \\ & & & \vdots \\ & & & 1 \end{pmatrix}}_{=D} \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \vdots \\ \alpha_n \end{pmatrix}}_{=\alpha} + \underbrace{\begin{pmatrix} x'_{1,1} \\ \vdots \\ x'_{1,T} \\ x'_{2,1} \\ \vdots \\ x'_{2,T} \\ \vdots \\ \vdots \\ x'_{n,1} \\ \vdots \\ x'_{n,T} \end{pmatrix}}_{=X} \beta + \underbrace{\begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,T} \\ \varepsilon_{2,1} \\ \vdots \\ \varepsilon_{2,T} \\ \vdots \\ \vdots \\ \varepsilon_{n,1} \\ \vdots \\ \varepsilon_{n,T} \end{pmatrix}}_{=\varepsilon}$$

$$M_D = I_{nT} - D(D'D)^{-1}D' = I_n \otimes M_T, \quad M_T = I_T - \frac{1}{T}S_T S_T' \quad S_T = (1, \ldots, 1)'$$

## Heteroscedasticity in the RE Model

Heteroscedasticity in both $\varepsilon_{it}$ and $u_i$ ?

$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it}, \; E[\varepsilon_{it}^2 \mid \mathbf{X}_i] = \sigma_{\varepsilon,i}^2, \; E[u_i^2 \mid \mathbf{X}_i] = \sigma_{u,i}^2$

OLS

$$\mathbf{b} = \left[\Sigma_{i=1}^N \mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[\Sigma_{i=1}^N \mathbf{X}_i'\mathbf{y}_i\right]$$

$$\text{Var}[\mathbf{b} \mid \mathbf{X}] = \frac{\mathbf{1}}{\Sigma_{i=1}^N T_i}\left[\frac{\mathbf{X}'\mathbf{X}}{\Sigma_{i=1}^N T_i}\right]^{-1}\frac{\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}}{\Sigma_{i=1}^N T_i}\left[\frac{\mathbf{X}'\mathbf{X}}{\Sigma_{i=1}^N T_i}\right]^{-1}$$

$\boldsymbol{\Omega} = \text{diag}[\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \ldots \boldsymbol{\Omega}_N]$ Each block is $T_i x T_i$

$\boldsymbol{\Omega}_i = \sigma_{\varepsilon,i}^2 \mathbf{I} + \sigma_{u,i}^2 \mathbf{ii}'$

## Autocorrelation

- Source?
- Already present in RE model – equicorrelated.
- Models:
  - Autoregressive: $\varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + v_{it}$ – how to interpret
  - Unrestricted: (Already considered)
- Estimation requires an estimate of ρ

$$\hat{\rho} = \frac{1}{N}\Sigma_{i=1}^N\left(\frac{\Sigma_{t=2}^{T_i} e_{i,t}e_{i,t-1}}{\Sigma_{t=1}^{T_i} e_{i,t}^2}\right) = \frac{1}{N}\Sigma_{i=1}^N \hat{\rho}_i$$

using LSDV residuals in both RE and FE cases

## FGLS – Fixed Effects

$y_{i,t} = \alpha_i + \mathbf{x}_{i,t}'\boldsymbol{\beta} + \varepsilon_{i,t}, \; \varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + v_{it}$

$y_{i,t} - \rho y_{i,t-1} = \alpha_i(1-\rho) + (\mathbf{x}_{i,t}' - \rho\mathbf{x}_{i,t-1}')\boldsymbol{\beta} + v_{i,t}$

$y_{i,t}{*} = \alpha_i{*} + \mathbf{x}_{i,t}'{*}\,\boldsymbol{\beta} + v_{i,t}$

Using $\hat{\rho}$ in LSDV estimation estimates

$\alpha_i{*} = \alpha_i(1-\rho)$,

$\boldsymbol{\beta}$,

$\sigma_v^2 = \sigma_\varepsilon^2(1-\rho^2)$

Estimate $\alpha_i$ with $a_i{*}/(1-\hat{\rho})$,

Estimate $\sigma_\varepsilon^2$ with $\hat{\sigma}_v^2/(1-\hat{\rho}^2)$

# FGLS – Random Effects

$y_{i,t} = \mathbf{x}'_{i,t}\boldsymbol{\beta} + u_i + \varepsilon_{i,t},\ \varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + v_{it}$

$y_{i,t} - \rho y_{i,t-1} = u_i(1-\rho) + (\mathbf{x}'_{i,t} - \rho\mathbf{x}'_{i,t-1})\boldsymbol{\beta} + v_{i,t}$

$y_{i,t}{*} = \mathbf{x}'_{i,t}{*}\boldsymbol{\beta} + u_i{*} + v_{i,t}$

(1) Step 1. Transform data using $\hat{\rho}$ to partial deviations

(2) Residuals from transformed data to estimate variances

Step 2 estimator of $\sigma_v^2$ using LSDV residuals

Step 2 estimator $\sigma_v^2 + \sigma_u^2(1-\rho^2) = \sigma_v^2 + \sigma_u^2{*}$

(3) Apply FGLS to transformed data to estimate $\boldsymbol{\beta}$ and asymptotic covariance matrix

(4) Estimates of $\sigma_u^2$, $\sigma_\varepsilon^2$ can be recovered from earlier results.

# Covariance Structures

- Model Structure
- Seemingly Unrelated Regressions
- OLS Estimation and 'Panel Corrected Standard Errors'
- GLS and FGLS Estimation – problem of too many variance parameters estimated

# Covariance Structures

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

$E[\varepsilon_{it} \mid \mathbf{X}] = 0$

$E[\varepsilon_{it}\varepsilon_{js} \mid \mathbf{X}] = 0$ if $t \neq s$, $\sigma_{ij}$ if $t = s$.

**Note, covariances across observations instead of across time.**

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \ldots \\ y_{1T} \\ y_{21} \\ y_{22} \\ \ldots \\ y_{2T} \end{pmatrix} = \begin{bmatrix} \mathbf{x}'_{11} \\ \mathbf{x}'_{12} \\ \ldots \\ \mathbf{x}'_{1T} \\ \mathbf{x}'_{21} \\ \mathbf{x}'_{22} \\ \ldots \\ \mathbf{x}'_{2T} \end{bmatrix}\boldsymbol{\beta} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \ldots \\ \varepsilon_{1T} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \ldots \\ \varepsilon_{2T} \end{pmatrix}, \mathrm{Var}[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \begin{bmatrix} \begin{pmatrix} \sigma_{11} & 0 & \ldots & 0 \\ 0 & \sigma_{11} & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma_{11} \end{pmatrix} & \begin{pmatrix} \sigma_{21} & 0 & \ldots & 0 \\ 0 & \sigma_{21} & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma_{21} \end{pmatrix} \\ \begin{pmatrix} \sigma_{21} & 0 & \ldots & 0 \\ 0 & \sigma_{21} & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma_{21} \end{pmatrix} & \begin{pmatrix} \sigma_{22} & 0 & \ldots & 0 \\ 0 & \sigma_{22} & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma_{22} \end{pmatrix} \end{bmatrix}$$

## Generalized Regression

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \cdots \\ \mathbf{y}_N \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \cdots \\ \mathbf{X}_N \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \cdots \\ \boldsymbol{\varepsilon}_N \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \ E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0},\ E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \boldsymbol{\Sigma} \otimes \mathbf{I}$$

## OLS Estimation

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$
$$= \left[\Sigma_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[\Sigma_{i=1}^{N} \mathbf{X}_i'\mathbf{y}_i\right]$$

The usual results for applying OLS in a GR model apply

$$\text{Var}[\mathbf{b}|\mathbf{X}] = \left[\Sigma_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i\right]^{-1}[\mathbf{X}'(\boldsymbol{\Sigma} \otimes \mathbf{I})\mathbf{X}]\left[\Sigma_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i\right]^{-1}$$
$$= \left[\Sigma_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left(\Sigma_{i=1}^{N}\Sigma_{j=1}^{N} \sigma_{ij}\mathbf{X}_i'\mathbf{X}_j\right)\left[\Sigma_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i\right]^{-1}$$

Estimating $\boldsymbol{\Sigma}$

$$\hat{\sigma}_{ij} = \frac{\Sigma_{t=1}^{T} e_{it}e_{jt}}{T} = s_{ij} \quad [s_{ij}] = \mathbf{S} = \hat{\boldsymbol{\Sigma}}$$

## Panel Corrected Standard Errors

Est.Asy.Var$[\mathbf{b}|\mathbf{X}]=$

$$\left[\Sigma_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left(\Sigma_{i=1}^{N}\Sigma_{j=1}^{N} s_{ij}\mathbf{X}_i'\mathbf{X}_j\right)\left[\Sigma_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i\right]^{-1}$$

## Statistical Inference with Panel Data

With panel data the errors are potentially
- Serially correlated (autocorrelation) (i.e., correlated over t for given i)
- Heteroscedastic (i.e., not constant over i)

Ignoring this leads to under-estimated S.E.s
- With RE- and FE-models the problem is alleviated a little
  - E.g., FE allows for equi-correlated errors over t

Solution I: Assume a particular error structure
- In Stata with
  - `xtgee`: generalized linear models with unit-specific correlation structure
  - `xtregar`: panel regression with first-order autoregressive error term
- Drawback: results heavily depend on the assumptions made
- And: `xtgee` estimates pooled or RE-models!
  - Nevertheless, many authors use these methods and are proud that they use such sophisticated statistical models. They are not aware that with these methods they do not make use of the main advantage of panel data (within estimation)!

## Stata

## xtregar

- Linear model with first order autoregressive error term -xtregar-:

$$Y_{it} = \alpha + X_{it}^1 * \beta_1 + ... + X_{it}^K * \beta_K + \mu_i + \nu_{it}$$

$$\nu_{it} = \rho * \nu_{it-1} + \eta_{it} \quad ; \quad \eta \ is \ iid(0, \sigma_\eta^2)$$

Some Relevant Assumptions
- Fixed effects: The non-observable individual effects ($\mu_i$) are represented by fixed parameters and may be correlated with the covariates in X.
- Random effects: The non-observable individual effects are assumed to be independent of the idiosyncratic error term and they are also independent of the covariates in X. The $\mu_i$ are iid($0,\sigma_\mu^2$)

## xtregar: Marriage Premium example

xtregar wage marr, fe

Ver Marriagepremium3.do

## xtregar: Country example

- Fit model accounting for autocorrelation

```
.   xtregar lconsumo lpib lirate, fe
FE (within) regression with AR(1) disturbances  Number of obs      =      2779
Group variable: country                         Number of groups   =       122
R-sq:  within  = 0.9631                         Obs per group: min =        10
       between = 0.9941                                        avg =      22.8
       overall = 0.9930                                        max =        30
                                                F(2,2655)          =  34634.76
corr(u_i, Xb)  = -0.2531                        Prob > F           =    0.0000

    lconsumo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpib |   .9887413   .0037579   263.11   0.000     .9813726      .99611
      lirate |   -.000825   .0021888    -0.38   0.706    -.0051168    .0034668
       _cons |   .0431147    .015465     2.79   0.005       .01279    .0734394
-------------+----------------------------------------------------------------
      rho_ar |  .82953965
     sigma_u |  .15647357
     sigma_e |  .04443063
     rho_fov |  .92538831   (fraction of variance because of u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(121,2655) =     9.24            Prob > F = 0.0000
```

## xtgls

- Feasible Generalized Least Squares (-xtgls-):

$Y_{it} = \alpha + X_{it}^{1} * \beta_1 + ... + X_{it}^{K} * \beta_K + \epsilon_{it}$

*Where the variance matrix of the disturbances would be :*

$$E[\epsilon\epsilon] = \Omega = \begin{pmatrix} \sigma_{11}\Omega_{11} & \sigma_{12}\Omega_{12} & \cdots & \sigma_{1m}\Omega_{1m} \\ \sigma_{21}\Omega_{21} & \sigma_{22}\Omega_{22} & \cdots & \sigma_{2m}\Omega_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1}\Omega_{m1} & \sigma_{m2}\Omega_{m2} & \cdots & \sigma_{mm}\Omega_{mm} \end{pmatrix}$$

Different variance-covariance structures

- Heteroskedasticity across panels
- Correlation across panels
- Autocorrelation within panels

# xtgls

- Feasible Generalized Least Squares -xtgls-:

$Y_{it} = \alpha + X_{it}^1 * \beta_1 + ... + X_{it}^K * \beta_K + \epsilon_{it}$

  - Heteroskedasticity across panels -xtgls,panels(heteroskedastic)-:

$$E[\epsilon\epsilon] = \Omega = \begin{pmatrix} \sigma_1 I & 0 & \cdots & 0 \\ 0 & \sigma_2 I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m I \end{pmatrix}$$

# xtgls

- Feasible Generalized Least Squares -xtgls-:

$Y_{it} = \alpha + X_{it}^1 * \beta_1 + ... + X_{it}^K * \beta_K + \epsilon_{it}$

$\epsilon_{it} = \rho * \epsilon_{it-1} + \eta_{it} \quad ; \quad \eta \ \ is \ \ iid(0, \sigma_\eta^2)$

  - Heteroskedasticity across panels and autocorrelation within panels -xtgls,panels(heteroskedastic) corr(psar1)-:

$$E[\epsilon\epsilon] = \Omega = \begin{pmatrix} \sigma_1 \Omega_{11} & 0 & \cdots & 0 \\ 0 & \sigma_2 \Omega_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m \Omega_{mm} \end{pmatrix}$$

# xtgls: Country Example

- Fit model accounting for heteroskedasticity

```
.  xtgls lconsumo lpib lirate,panels(heterosk) nolog
Cross-sectional time-series FGLS regression
Coefficients:  generalized least squares
Panels:        heteroskedastic
Correlation:   no autocorrelation
Estimated covariances      =       122       Number of obs       =      2901
Estimated autocorrelations =         0       Number of groups    =       122
Estimated coefficients     =         3       Obs per group: min  =        11
                                                            avg  =  23.77869
                                                            max  =        31
                                             Wald chi2(2)        =   1352635
                                             Prob > chi2         =    0.0000

    lconsumo |     Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpib |  .9720376    .000853  1139.53   0.000     .9703657    .9737094
      lirate |  .0193624   .0013921    13.91   0.000      .016634    .0220908
       _cons |  .4192768   .0215707    19.44   0.000      .376999    .4615545

.
```

# xtgls: Country Example

- Fit model accounting for autocorrelation and heteroskedasticity

```
.  xtgls lconsumo lpib lirate,panels(heterosk) corr(psar1) nolog force
Cross-sectional time-series FGLS regression
Coefficients:  generalized least squares
Panels:        heteroskedastic
Correlation:   panel-specific AR(1)
Estimated covariances      =       122          Number of obs      =      2901
Estimated autocorrelations =       122          Number of groups   =       122
Estimated coefficients     =         3          Obs per group: min =        11
                                                               avg =  23.77869
                                                               max =        31
                                                Wald chi2(2)       = 255484.88
                                                Prob > chi2        =    0.0000
```

| lconsumo | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lpib | .957914 | .0019214 | 498.54 | 0.000 | .9541481 | .96168 |
| lirate | -.0009035 | .0009393 | -0.96 | 0.336 | -.0027444 | .0009375 |
| _cons | .7854506 | .0472898 | 16.61 | 0.000 | .6927643 | .8781369 |

.

# Example 2: Investment Demand

- A classical panel data model of investment demand (Greene [2008], pp.250-252, Grunfeld and Griliches [1960]) is defined by: $I_{it} = \alpha_i + \beta F_{it} + \gamma C_{it} + \varepsilon_{it}$
- where
  i = 10 firms: GM, CH, GE, WE, US, AF, DM, GY, UN, IBM.
  t = 20 years: 1935-1954.
  $I_{it}$ = Gross investment.
  $F_{it}$ = Market value.
  $C_{it}$ = Value of the stock of plant and equipment.
  $e_{it}$ = Error term.

- Example1_1.do

# xtreg pa

Population-averaged model, allows within-group correlation structures to be specified for the panels

```
•xtreg i f c, pa
•xtreg i f c, pa vce(robust)
•xtreg i f c, pa vce(robust) corr(exchangeable)
•xtreg i f c, pa vce(robust) corr(stationary 1)
•xtreg i f c, pa vce(robust) corr(nonstationary 1)
•xtreg i f c, pa vce(robust) corr(ar 1)
•xtreg i f c, pa vce(robust) corr(unstructured)
```

```
Example1_2.do
```

## Pooled Regression GLS

- Examples of Time Series Correlation
  - Equal-Correlation $$\rho_{ts} = \begin{cases} 1 & if\ t = s \\ \rho & if\ t \neq s \end{cases}$$
  - AR(1) $$\rho_{ts} = \begin{cases} 1 & if\ t = s \\ \rho^{|t-s|} & if\ t \neq s \end{cases}$$
  - Stationary(1) $$\rho_{ts} = \begin{cases} 1 & if\ t = s \\ \rho & if\ |t-s| = 1 \\ 0 & otherwise \end{cases}$$
  - Nonstationary(1) $$\rho_{ts} = \begin{cases} 1 & if\ t = s \\ \rho_{ts} & if\ |t-s| = 1, \rho_{ts} = \rho_{st} \\ 0 & otherwise \end{cases}$$

## Pooled Regression GLS

* pooled regression: GLS

```
xtgls i f c, panels(hetero)            // hetero. across panels
xtgls i f c, panels(hetero) corr(ar1)  // hetero. and AR(1) within panels
xtgls i f c, panels(hetero) corr(psar1) // hetero. and panel specific AR(1)
```

* hetero. and cross-sec. correlation

```
xtgls i f c, panels(correlated)
matrix list e(Sigma)
xtgls i f c, panels(correlated) corr(ar1)   // with AR(1) within panels matrix list
e(Sigma)
xtgls i f c, panels(correlated) corr(psar1) // with panel specific AR(1)
matrix list e(Sigma)
```

* panel-corrected s.e.

```
xtpcse i f c, correlation(ar1)
xtpcse i f c, correlation(psar1)


example1_4.do
Xtpce and xtgls are more suited than xtgee for long panels.
```

## xtoverid: Robust Hausman Test

Abdata.dat

```
*(Balanced panel)
 xtreg n w k if year>=1978 & year<=1982, re
xtoverid
est store re

xtreg n w k if year>=1978 & year<=1982, fe
est store fe

*(In homoskedastic balanced panel case, Hausman test using sigma from
FE estimation...)
 hausman fe re, sigmaless


*(Artificial regression overid statistic readily extends to non-
homoskedastic case)
xtreg n w k, re cluster(id)
xtoverid
```