Silvia Mendolia

Labour market, health and well-being: economic analysis using panel data

Plan for today

- Other types of fixed effects
- Random effects model
- Test and diagnostics: RE or FE?
- Correlated Random Effects
- Limitations of panel data estimation
- How to interpret your results and present possible limitations

Other types of fixed effects

- Time FE (use time dummies to understand if there is a specific effect of a certain point in time)

- Country/Region FE (use country or region dummies to control for the time-invariant characteristics of a certain geographical area)

Other types of fixed effects

- School/Class FE (to capture the common factors in a school/class, eg teachers' attitudes, environment, prejudices etc)
- Sibling FE (to capture common characteristics of kids raised in the same family)

Time FE

- To see if time FE (eg year dummies) are needed we may use testparm after running the command in Stata
- This is a joint test to see if the dummies for all years are equal to 0
- xtreg sclfsato age_dv i.wave i.sex_dv married separated widow i.hiqual_dv self_employed unemployed outlforce, fe

```
. testparm i.wave

 ( 1)  2.wave = 0
 ( 2)  3.wave = 0
 ( 3)  4.wave = 0
 ( 4)  5.wave = 0
 ( 5)  6.wave = 0
 ( 6)  7.wave = 0
 ( 7)  8.wave = 0
 ( 8)  9.wave = 0

       F(  8,140146) =   95.68
            Prob > F =    0.0000
```

We reject the null hp that the coefficients for all years are jointly equal to 0 and so we need time FE

Random effects estimator

- **FE estimator wipes out the between variation** (variation across cross sections)
- If we believe that the a_i 's are not correlated with the regressors, we should resort to more efficient estimators
- RE estimator could be a feasible estimator

Random effects estimator

$$Y_{it} = \alpha + x_{it}\beta + a_i + \varepsilon_{it}$$

- Suppose a_i is purely random with mean 0 and variance σ^2_a
- In other words, we are assuming a_i is **uncorrelated** with the regressors

- This is a **very strong assumption** (even stronger than what needed for FE)

Random effects estimator

- The RE estimator is a **weighted average of a within and between model**

$$
\begin{aligned}
(y_{it} - \overline{y_i}) + (1 - \hat{\theta}_i)\overline{y}_i \quad = \quad & (x_{it} - \overline{x}_i)\beta + (\epsilon_{it} - \overline{\epsilon}_i) \\
& + (1 - \hat{\theta}_i)(\alpha + \overline{x}_i\beta + a_i + \overline{\epsilon}_i)
\end{aligned}
$$

- **In Stata, we use xtreg, re**

Random effects estimator

$$Y_{it} = \alpha + x_{it}\beta + a_i + \varepsilon_{it}$$

- The logic of RE is that, unlike FE, the variation across individuals is assumed to be random and uncorrelated with the independent variables in the model
- "…the crucial distinction between fixed and random effects is **whether the unobserved individual effect embodies elements that are correlated with the regressors** in the model, not whether these effects are stochastic or not" [Green, 2008, p.183]

Random effects estimator

$$Y_{it} = \alpha + x_{it}\beta + a_i + \varepsilon_{it}$$

- An advantage of RE is that **we can include time-constant independent variables** (eg gender and in many cases education)
- In the RE model, you need to include individual characteristics that may or may not influence the predictor variables.
- However, some variables may not be available
- Omitted variable bias in the model?

Comparing RE and FE

- Let’s use the last example we made and compare RE and FE estimates
- Use longitudinal_td.dta
- xtreg sclfsato age_dv i.sex_dv married separated widow i.hiqual_dv self_employed unemployed outlforce, fe
- estimates store FE
- xtreg sclfsato age_dv i.sex_dv married separated widow i.hiqual_dv self_employed unemployed outlforce, re
- estimates store RE
- estimates table FE RE, b se

Comparing RE and FE

```
. estimates table FE RE, b se
```

Variable	FE	RE
age_dv	-.00642314	.00616264
	.00122866	.00041506
sex_dv		
Female	(omitted)	.0415845
		.01256114
married	.16085272	.22832874
	.02412544	.01476111
separated	-.02995826	-.20300746
	.03080957	.02036037
widow	-.04190701	.07657132
	.04028489	.02542923
hiqual_dv		
Other hig..	.1084236	-.06140337
	.05597975	.02010287
A level etc	.17343366	-.07522068
	.04605859	.01781304
GCSE etc	.18094291	-.1816244
	.06000858	.01796659
Other qual	.08059962	-.22135884
	.08320697	.02208995
No qual	.08769546	-.25807731
	.088886	.0206778
self_emplo~d	.02593077	.00009619
	.01978107	.01615603
unemployed	-.21119275	-.34781119
	.02047336	.01791513
outlforce	.02052477	.00540641
	.013382	.01055973
_cons	5.3308938	4.838006
	.08000153	.02384111

legend: b/se

RE or FE

- If the individual FE is correlated with the regressors, we must use FE (the only consistent estimator)
- If individual FE is uncorrelated with the regressors, both RE and Fe are consistent but RE is more efficient
- This is because RE uses between and within variation

RE or FE

- **The Hausman test can be used to choose RE or FE**
- It starts with the idea that both estimators are consistent under the null hypothesis of no correlation
- Under the null hp the estimates of β should not differ systematically
- Use **hausman** in Stata to run the test

RE or FE

- The intuition is: use RE only if Hausman test says it's ok!
- FE is consistent. If RE does not differ too much then we can use RE
- If you **reject H0**, use FE
- If you are **not** able to reject H0, use RE

RE or FE

```
hausman FE RE, sigmamore

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       FE           RE         Difference          S.E.
-------------+----------------------------------------------------------------
      age_dv |   -.0064231     .0061626       -.0125858        .0011554
     married |    .1608527     .2283287        -.067476        .0190593
   separated |   -.0299583    -.2030075        .1730492        .0230919
       widow |    -.041907     .0765713       -.1184783        .0312049
   hiqual_dv |
          2  |    .1084236    -.0614034         .169827        .0521998
          3  |    .1734337    -.0752207        .2486543        .0424364
          4  |    .1809429    -.1816244        .3625673        .0572078
          5  |    .0805996    -.2213588        .3019585        .0801552
          9  |    .0876955    -.2580773        .3457728        .0863776
elf_emplo~d  |    .0259308     .0000962        .0258346        .0113875
  unemployed |   -.2111927    -.3478112        .1366184        .0098776
   outlforce |    .0205248     .0054064        .0151184        .0082034
------------------------------------------------------------------------------
                          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(12) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                           =      606.75
                Prob>chi2 =      0.0000
```

If this is <0.05 use FE

RE or FE

- Given the test, we reject the null hypothesis of no correlation
- **We conclude that the estimators are different and we must use FE**

Limitations of panel data analysis

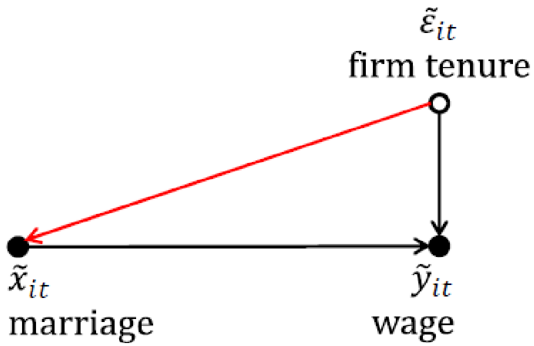
- **FE estimates are biased under endogeneity**

$E(x_{is}\ \varepsilon_{it}) \neq 0$ for s, t=1, ….T

- Endogeneity can have several sources
  - Unobserved time-varying confounders
  - Measurement error (errors in reporting x)
  - Y affects X (reverse causality)
  - Endogenous selection bias
- When using FE, we need to argue that none of these sources of endogeneity is problematic

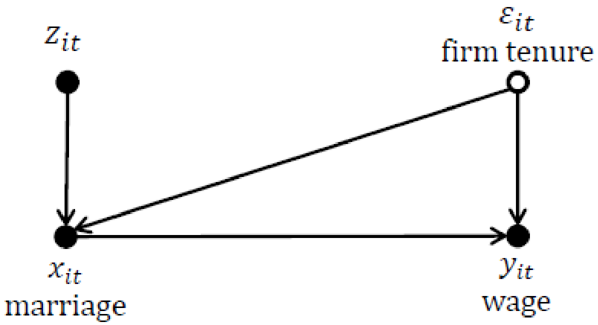
Sources of endogeneity

- **Unobserved time-varying variables**
  - eg: is there a marriage premium (does marriage increase wages?)
  - Strict exogeneity assumption: firm tenure does not affect chances of marriage
  - However: longer firm tenure may indicate a more secure job and this in turn may affect chances of marriage



Sources of endogeneity

- How to solve this problem?
  - Instrumental Variables can be used
  - Stata routine is xtivreg
  - In this example, we need to find a variable (instrument) that affects marriage but does not directly affect wage



Sources of endogeneity

- **Measurement error**
  - If the model includes only one variable, measurement error generally produce "**attenuation bias**"
  - With more variables, the direction of the bias is unknown
  - When using FD or FE estimator, the measurement error problem may be amplified
  - Taking the difference of two unreliable measures can produce an even more unreliable measure
  - However, FD or FE reduces the unobserved heterogeneity problem and simulation studies show this bias dominates

Sources of endogeneity

- Measurement error
  - People respond to the same question in different ways
  - Example: life satisfaction. Different people mean something different when they say they are completely satisfied with their life
  - The problem is amplified with ranking scales (eg rank your life satisfaction from 1 to 7)
  - If this type of error is related to time-invariant characteristics (eg personality traits) FE estimates will not be biased

Sources of endogeneity

- Reverse causality
  - Reverse causality (Y affects X)
  - Reciprocal causality (Y affects X and X affects Y)

- **Any econometric model is biased under reciprocal causality!**

- We can mitigate the risk of reverse causality by using panel data and looking at the timing of the events

Sources of endogeneity

- Selection into treatment
  - Treatment X affects outcome Y
  - People with certain levels of Y are more likely to get X
  - This means some people may select into treatment
  - FE can help to deal with selection
    - Example: people living with a partner from a different ethnic group or religion are less likely to have stereotypes against immigrants
    - BUT
    - People with less stereotypes are more likely to marry someone with a different background
  - FE can help in this context (even if it may not be perfect)

Panel attrition

- Panel attrition can be a threat to panel analysis
- However, the main problem comes from attrition deriving from unobservable variables

- Attrition deriving from time-invariant variables is not a problem for FE estimation
- Example: people with certain personality traits are more likely to leave the survey

Correlation over time

- Panel variables can be correlated over time
  - Outcomes depend on characteristics that are fixed over time
    - FE estimates are unbiased
  - Past outcomes affect future outcomes
    - We can use models with lagged dependent variable as a regressor (dynamic panel models)
    - However, FE is not feasible here (strict exogeneity assumption is violated)
    - We can use IV estimation

How to interpret your results – Caution!

- No estimation method will ever be perfect!
- If you have doubts about the strict exogeneity assumption, **explain the *caveats* carefully after your present the results**
- Use caution in interpreting the results as causal if you have doubts about the assumption
- It is better to show caution than to look overconfident!

How to interpret your results – Caution!

- **It is very hard to rule out reverse causality by simply using panel data**
- The timing of events is not necessarily helpful, because there may be anticipation effects and measurement error
- Sometimes it is hard to know whether X or Y changed first
- **Be upfront and explain the potential limitations of your work!**