

Topic 5 Simulation: Fixed Effects Models and an introduction to Random Effects

We are commonly concerned with unobserved effects being correlated with observed explanatory variables as we cannot control for all potential factors in a regression and this issue biases our results. Thus the issue is that our unobserved factors are correlated with our explanatory variables x as well as with our error term u .

For example we might be concerned that intelligence is correlated with years of schooling as well as future expected earnings. However, fortunately, intelligence is thought of as a (relatively) time constant factor. Therefore, if we remove time constant factors we might be able to approximate the returns to education. (This is assuming the returns to years of education is constant. If it is a function of intelligence then we are going to need to think about being a little cleverer than this.)

In order to remove this potentially biasing effect we could think about taking measurements over multiple periods for the same individual inquiring about their level of education and their earnings. If we observe that within the same individual, removing the time constant effects (intelligence) there is still a relationship between education and wage, then we may feel more secure in our hypothesis that education leads to higher earnings.

In the lectures we talked through the fixed effects method for removing this time invariant factors (intelligence in this case), to allow for accurate estimation. This model takes the time average for an individual (i) away from their current value to allow for this unobserved time invariant term (α_i) to be difference away:

$$\begin{aligned} Wage_{it} - \overline{Wage}_i &= \beta_0 + \beta_1(Education\ Level_{it} - \overline{Education\ Level}_i) + \\ &\beta_2(Experience_{it} - \overline{Experience}_i) + (\alpha_i - \alpha_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \end{aligned}$$

As always let's start by simulating some data, this will be slightly more complex than normal as we need a panel rather than just a cross section but I think it will be easier than the more complex IV material we worked through last week. Let's see!

To start with here I create the number of observations and then create a unique identifier to show which observations belong to which individuals. We have created our first variable c which will be our time constant heterogeneity. I have also included some labels this week so we can keep track of what our variables are in the dataset a little easier. We should now have observation of c per individual. The next step is to create numerous observations for each individual and to create a year variable so we can keep track of them. If you are interested in the commands there are two important lines here. The first is the `expand 5` line, here we are expanding out the dataset so rather than one observation per person we should now have five. The second command which is very useful in a wide variety of applications is the `"bysort"` command. This sorts by `id`, for for year individual `id` number runs the command after the colon. So in this case the `bysort` command runs the `gen year= _n` command on subsets of the dataset one by one, with each subset being those with the same `id` here. Now is a good point to browse the dataset. We should have 1000 observations, with 200 unique `id`'s, each repeated 5 times. The variable c should be the same in each time period for an individual as it is time invariant. As well as browsing the data we can quickly check the data by tabulating the year variable which I have included below.

```

clear
set obs 200
set seed 101
gen c = rnormal()
label var c "Time Constant Heterogeneity (individual specific)"
gen id = _n
label var id "Individual specific ID"
expand 5
bysort id: gen year=_n
label var year "Year of observation"
tab year

```

What's next? We only have one variable at the moment, our time invariant constant so we should start adding other variables so we can run our first regression on our newly formed panel.

```

gen x = rnormal()+c
label var x "explanatory variable X (with time constant and time varying components)"
gen u = 3*rnormal()+3*c
label var u "Error term (correlated with unobservables c)"
gen y = x + u
label var y "Outcome variable"
reg y x

```

Our results here (Figure 1) seem to show that OLS is biased, we should be getting an estimate of the x coefficient of around one but it is over two. The unobserved time invariant variable c is related to both our x variable and our error term and thus is biasing our result.

Figure 1: Regression Output for OLS, $y=x$

Source	SS	df	MS	Number of obs	=	1,000
Model	10106.9199	1	10106.9199	F(1, 998)	=	830.11
Residual	12151.1074	998	12.1754584	Prob > F	=	0.0000
				R-squared	=	0.4541
				Adj R-squared	=	0.4535
Total	22258.0273	999	22.2803076	Root MSE	=	3.4893

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	2.394168	.0830975	28.81	0.000	2.231102	2.557234
_cons	-.1028941	.1103854	-0.93	0.351	-.3195083	.1137201

Let's look at running a panel data model on this rather than just a pooled OLS technique. The first thing to do is to tell STATA this is panel data using the xtset command, outlining that year is the time element of the panel and id is the individual identifier. Let's now look at the different ways we can run a fixed effects regression. The first is using xtreg y x, fe. This should be exactly the same as running a standard regression with a dummy variable in for each individual in the

dataset which I have replicated with the command below it. I have only included the results from the first one here as the second one produces a lot of output!

```
xtset id year
xtreg y x, fe
```

As we can see in Figure 2 below the FE regression has removed the bias and has now provided a much more accurate estimate of the coefficient on x, with the confidence interval including the true value of x. The fixed effect estimator is unbiased because it successfully eliminates the correlation between the time constant correlation between the x and the error u.

Figure 2: Regression Output for FE y=x

```
Fixed-effects (within) regression               Number of obs   =       1,000
Group variable: id                             Number of groups =        200

R-sq:                                           Obs per group:
    within = 0.1051                             min =           5
    between = 0.8164                             avg =          5.0
    overall = 0.4541                             max =           5

corr(u_i, Xb) = 0.6236                        F(1,799)        =       93.87
                                              Prob > F         =       0.0000
```

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	1.0812	.1115958	9.69	0.000	.8621444	1.300256
_cons	-.0542316	.0961552	-0.56	0.573	-.2429781	.134515
sigma_u	2.8029021					
sigma_e	3.0378789					
rho	.45983464	(fraction of variance due to u_i)				

F test that all u_i=0: F(199, 799) = 2.60 Prob > F = 0.0000

What about random effects?

We have detailed previously that we use FE when there is some relationship between our α_i term and one of our independent variables. FE let's us remove this by differencing it away.

Does this always happen or do we sometimes have an unobserved heterogeneity that is not associated with the error term? Well actually in some situations this is actually the case. What should we do then? Well one answer might be to use pooled OLS. Pooled OLS here will be consistent, and we could still run a FE model if we want. FE is quite a strict model and we do lose quite a lot of information so it might not be our best bet.

Pooled OLS then? We are still going to have what we call a composite error term in this model if we run OLS as we don't observe either α_i or ε_{it} . Thus we will have a composite error term, let's call it η_{it} which is equal to $\alpha_i + \varepsilon_{it}$. Why is this important? One of the OLS assumptions is that our error terms are not serially correlated, i.e. η_{it} does not have a relationship to η_{is} where s just means any other time period apart from t . This is normally okay if we are just dealing with ε_{it} , but now we have η_{it} this causes a few issues. We need $\alpha_i + \varepsilon_{it}$ to not relate to $\alpha_i + \varepsilon_{is}$. We may be happy that $\alpha_i + \varepsilon_{it}$ don't relate, and that $\alpha_i + \varepsilon_{is}$, and even that $\varepsilon_{it} + \varepsilon_{is}$. However, our time invariant characteristic α_i will be the same in all time periods and thus part of our composite error will be serially correlated. In this case random effects is our best bet as it will be more efficient than pooled OLS and FE!

Let's return to our example from the lecture where house prices (HP) are determined by crime in the area (CRIME), unemployment in the area (UNEMP), an unobserved time invariant regional characteristic α_i and an error term.

$$HP_{it} = \beta_0 + \beta_1 CRIME_{it} + \beta_2 UNEMP_{it} + \alpha_i + \varepsilon_{it}$$

We are assuming that α_i and all our independent variables are not related. In this case we tend to write RE as we can see below with the original equation, minus its time average (the same as FE), but this time the time average is multiplied by a term λ . Thus every term has its original minus the time average multiplied by λ . I have written the error here in its composite form that we introduced earlier.

$$HP_{it} - \lambda \overline{HP_t} = \beta_0(1 - \lambda) + \beta_1(CRIME_{it} - \lambda \overline{CRIME_t}) + \beta_2(UNEMP_{it} - \lambda \overline{UNEMP_t}) + \eta_{it} - \lambda \overline{\eta_t}$$

Without saying what λ is we can say a few things about this equation. If $\lambda = 0$ then the time averaged part of all these equations will be equal to zero and our model just reduces back down to the previous pooled OLS equation.

$$HP_{it} = \beta_0 + \beta_1 CRIME_{it} + \beta_2 UNEMP_{it} + \eta_{it}$$

If $\lambda = 1$ then the above equation will be the same as FE, with just the time averaged part taken away from the original equation.

$$HP_{it} - \overline{HP_t} = \beta_0 + \beta_1(CRIME_{it} - \overline{CRIME_t}) + \beta_2(UNEMP_{it} - \overline{UNEMP_t}) + \eta_{it} - \overline{\eta_t}$$

What about values of λ between zero and one? This is the most common result, with λ usually being between zero and one. If this is the case the RE is not equal to FE or OLS. It is better than both and should give a lower standard error!

RE measures the impact of the α_i on the equation. As the importance of this term fades to zero then the model tends towards OLS. As this α_i term trends towards infinity then the model tends towards FE. So a higher λ value means α_i is more important.

RE is often called a quasi-time demeaned model. Why is this? Because it does not fully time demean the variables, instead it takes off some portion (λ) of the time average. An added bonus of this model compared to FE is that it allows us to model time constant variables that would drop out of a typical FE regression.

The downsides? The assumption that α_i and all our independent variables are not related is not often met, and if it is not met then RE is not consistent. So if we use FE when we shouldn't it is

consistent (so as the sample tends to infinity the coefficient will tend towards it's true value), but it might not have the lowest standard errors. If we use RE when we shouldn't, it will not produce coefficient estimates that tend towards the true value as the sample size increases. There is a test to see if we should use FE or RE (the Hausman test) but this test pretty much always gives the result that FE should be used.

What would a simulation of RE data look like? The same as what we created above for FE but rather than have a relationship between x and c, c will only impact on the error term. See below:

```
gen x2 = rnormal()
label var x2 "explanatory variable X (time varying only)"

gen y2 = x2 + u
label var y2 "Outcome variable"
reg y2 x2
xtreg y2 x2, re
```

As we can see below pooled OLS does not do badly here, but RE does better.

Figure 3: Regression Output Pooled OLS y=1

Source	SS	df	MS	Number of obs	=	1,000
Model	1235.15712	1	1235.15712	F(1, 998)	=	79.19
Residual	15566.4628	998	15.5976581	Prob > F	=	0.0000
				R-squared	=	0.0735
				Adj R-squared	=	0.0726
Total	16801.6199	999	16.8184383	Root MSE	=	3.9494

y2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x2	1.108569	.1245751	8.90	0.000	.8641097	1.353028
_cons	-.0513224	.1248906	-0.41	0.681	-.2964008	.1937559

Figure 4: Regression Output RE y=x

```

Random-effects GLS regression              Number of obs   =      1,000
Group variable: id                        Number of groups  =       200

R-sq:                                     Obs per group:
    within = 0.1105                               min =          5
    between = 0.0377                              avg =         5.0
    overall = 0.0735                               max =          5

corr(u_i, X) = 0 (assumed)                  Wald chi2(1)     =     107.04
                                           Prob > chi2      =     0.0000

```

y2	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
x2	1.092415	.1055863	10.35	0.000	.8854698	1.29936
_cons	-.0513075	.203308	-0.25	0.801	-.4497838	.3471688
sigma_u	2.5355854					
sigma_e	3.0376922					
rho	.41063351	(fraction of variance due to u_i)				

Let's quickly look at running a manual version of the Hausman test¹ using some shortcuts from the Mundlak- Chamberlain approach.

This Mundlak-Chamberlain approach to running a FE regression is rather than include a dummy for each individual in the regression, to include a new variable which "controls" any time constant variation in x. We only have issues if there is some correlation between x and α_i and this can only be caused by the time invariant part of x.

This is easily done in STATA:

```

bysort id: egen x_bar = mean(x)
reg y x x_bar

```

The interesting thing about this is it allows us to test the idea that the time invariant part of x is important to the equation. If this x_bar term is significant it suggests that we need to use FE. If it is not significant it suggests we do not.

We should have two versions of x and y in our dataset now, one which was made so there was a relationship between α_i and x, and one where there wasn't. So let's run both and see if our Hausman test works!

```

bysort id: egen x_bar = mean(x)
reg y x x_bar
bysort id: egen x2_bar = mean(x2)
reg y2 x2 x2_bar

```

¹ There are a few versions of this test, the best one to look at is likely by Arellano (1993) {<http://ideas.repec.org/a/eee/econom/v59y1993i1-2p87-97.html>}.

In our first regression, where x is correlated to the unobserved time invariant heterogeneity, the x_bar coefficient is significant, indicating we should run FE. In Figure 6 where we know that the unobserved time invariant heterogeneity is not related to x , we see the x_bar coefficient is insignificant, suggesting that we can settle for either a pooled OLS or a RE regression.

Figure 5: Hausman Test on y and x

Linear regression	Number of obs	=	1,000
	F(2, 199)	=	463.36
	Prob > F	=	0.0000
	R-squared	=	0.5531
	Root MSE	=	3.1587

(Std. Err. adjusted for **200** clusters in id)

y	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
x	1.0812	.1111148	9.73	0.000	.8620864	1.300314
x_bar	2.264818	.1516391	14.94	0.000	1.965792	2.563844
$_cons$	-.1381725	.1138062	-1.21	0.226	-.3625934	.0862483

Figure 6: Hausman Test on $y2$ and $x2$

Linear regression	Number of obs	=	1,000
	F(2, 199)	=	54.21
	Prob > F	=	0.0000
	R-squared	=	0.0736
	Root MSE	=	3.9512

(Std. Err. adjusted for **200** clusters in id)

$y2$	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
$x2$	1.086408	.1094992	9.92	0.000	.8704798	1.302335
$x2_bar$	.0971373	.4622319	0.21	0.834	-.8143639	1.008639
$_cons$	-.0513918	.2031787	-0.25	0.801	-.4520513	.3492677

How bad is it if we run RE and we shouldn't have? If we go back to our previous FE example we can show this. The below code will set up our stereotypical FE situation, when the unobserved time invariant heterogeneity is correlated to the error term.

```

clear
set obs 200
set seed 101
gen c = rnormal()
label var c "Time Constant Heterogeniety (individual specific)"
gen id = _n
label var id "Individual specific ID"
expand 5
bysort id: gen year=_n
label var year "Year of observation"
tab year
gen x = rnormal()+c
label var x "explanatory variable X (with time constant and time varying components)"
gen u = 3*rnormal()+3*c
label var u "Error term (correlated with unobservables c)"
gen y = x + u
label var y "Outcome variable"
reg y x
xtset id year
xtreg y x, fe
xtreg y x, re

```

Running FE pulls out the correct coefficient on x, a value of 1, which is statistically significant. However if we look at our RE model in this situation as seen in Figure 8, the bias is very easy to see.

Figure 7: Regression Output FE, y=x

```

Fixed-effects (within) regression               Number of obs   =       1,000
Group variable: id                             Number of groups =        200

R-sq:                                           Obs per group:
    within = 0.1051                             min =           5
    between = 0.8164                             avg  =          5.0
    overall  = 0.4541                             max  =           5

corr(u_i, Xb) = 0.6236                        F(1,799)        =       93.87
                                           Prob > F         =       0.0000

```

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	1.0812	.1115958	9.69	0.000	.8621444	1.300256
_cons	-.0542316	.0961552	-0.56	0.573	-.2429781	.134515
sigma_u	2.8029021					
sigma_e	3.0378789					
rho	.45983464	(fraction of variance due to u_i)				

F test that all u_i=0: F(199, 799) = 2.60 Prob > F = 0.0000

We are now finding a coefficient on x of over twice it's actual size.

Figure 8: Regression Output RE, $y=x$

```

Random-effects GLS regression              Number of obs   =      1,000
Group variable: id                        Number of groups =       200

R-sq:                                     Obs per group:
    within = 0.1051                        min =          5
    between = 0.8164                       avg =         5.0
    overall = 0.4541                       max =          5

corr(u_i, X) = 0 (assumed)                Wald chi2(1)     =      640.37
                                           Prob > chi2      =      0.0000

```

y	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
x	2.201773	.0870079	25.31	0.000	2.031241	2.372306
_cons	-.0957634	.1251021	-0.77	0.444	-.340959	.1494322
sigma_u	.86834783					
sigma_e	3.0378789					
rho	.0755332	(fraction of variance due to u_i)				

Conclusions and take away comments

We have covered quite a lot in these notes, including introducing the all new random effects model. The main points to take away are why it is often a good idea to control for this unobserved time invariant heterogeneity with either FE or RE, depending on whether you believe this term is correlated to the error. If it is uncorrelated then RE is the best choice, if it is correlated FE is the best choice. Do note however, that if you use FE when RE would have been wiser the only downside is that the standard errors may not be as small as possible, it will still be unbiased. If you use RE when FE was the correct choice then your results will be biased. Thus FE is used a lot more in economics than RE. There are tests such as the Hausman to identify which is the most suitable model to run but if in doubt.... Try FE!

I think that is it for now..... we should be moving onto time series next week so watch this space!

David Morris