Recap of various contents

Federico Nutarelli

IV quick recap

If you have specific concerns, please send me an email.
I will provide you here some general tricks/guidance:

- The weak instrument bias tends to get worse as we add more weak instruments. In other words, the bias gets worse when there are many overidentifying restrictions (many instruments compared to endogenous regressors). By construction adding further instruments without predictive power reduces the value of F and the bias goes up (2SLS gets worse);
- If you have many potential IV, choose the best instrument and report the just identified model (weak instrument problem is less problematic in that case)

Let's move to R where we will pick up a famous problem from Angrist and Krueger(1991): on economic returns to schooling.
In practice it is always difficult to find convincing instruments (in particular satisfying the exclusion restriction).
The influential study of Angrist and Krueger 1991 used quarter of birth as an IV for schooling. Most states want student to enter school in the calendar year in which they turn 6.
Group A: children born in the 4th quarter enter school shortly before they turn 6;
Group B: children born in the 1st quarter enter school at a round age 6.5.
Law requires students to remain in school only until their 16th birthday.
Therefore A and B will have different ages when they start school and thus different lengths of schooling at the time they turn 16 when they can potentially drop out...

Clarifications on Panel data models and assumptions

In its more general form, the **true** (= unknown) dgp of a panel looks like this:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + z_i\gamma + \delta t + u_{it}, \quad i = 1, \ldots, N \quad t = \ldots, T \tag{1}$$

where, usually, $N >> T$.

Notation: The most common notations that you can find for panel data models include:

- The most specific notation: see Eq.(1)
- The broad notation for a general individual i: $y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{iT} \end{pmatrix}; X_i = \begin{pmatrix} x_{i,11} & \dots & x_{i,k1} \\ x_{i,12} & \dots & x_{i,k2} \\ x_{i,1s} & \dots & x_{i,ks} \\ x_{i,1T} & \dots & x_{i,kT} \end{pmatrix}$
  where k is the number of regressors. Notice that in some textbooks the pedex i is omitted. This means, implicitly, that the values of the X is different for each i. For instance:
  $$X_i = \begin{pmatrix} x_{11} & \dots & x_{k1} \\ x_{12} & \dots & x_{k2} \\ x_{1s} & \dots & x_{ks} \\ x_{1T} & \dots & x_{kT} \end{pmatrix}; \dots; X_j = \begin{pmatrix} w_{11} & \dots & w_{k1} \\ w_{12} & \dots & w_{k2} \\ w_{1s} & \dots & w_{ks} \\ w_{1T} & \dots & w_{kT} \end{pmatrix}$$
- The stacked notation:
  $$y = X\beta + \alpha + u$$
  In balanced panel data, y is an $NT \times 1$ vector as well as α. X is an $NT \times k$ matrix and $\beta \in k \times 1$.
  In unbalanced panel data we will have T_i for individual i rather than T since each individual is observed for a different amount of time. This means that, for instance $y \in \sum_i T_i \times 1$ (notice that if T_i is equal for all individulals, i.e. the panel is balanced, than $\sum_i T_i = NT$).

The 3 main panel data models are:

- Pooled OLS (POLS);
- Random Effects (RE);
- Fixed Effects (FE)
- First Difference (FD), less used

As you might know, Pooled OLS simply treat panels as large cross sections of individuals (so that individual i observed at time t is considered as a different observation from i observed at time s). FE and FD just differ from each other for the quantity that they subtract to y, x, u. Namely, FE subtract the temporal mean (mean taken over individuals) while FD subtracts the period before to each individual. While the formulas are equivalent (compact notation), i.e.:

$$\Delta y = \beta \Delta X + \Delta u$$

, if we decompose FE, we will have:

$$y_{it} - \overline{y}_i = (X_{it} - \overline{X}_i)\beta + u_{it} - \overline{u}_i$$

(where α_i is cancelled out since $\overline{\alpha}_i = \alpha_i$), and in FD,

$$y_{it} - y_{i,t-1} = (X_{it} - X_{i,t-1})\beta + u_{it} - u_{i,t-1}$$

, (where α_i is cancelled out since α_i is time independent, i.e. $\alpha_{i,t} = \alpha_{i,s} \quad \forall t, s$)

General intuition of the various models is given at the blackboard.

Conditions for Unbiasedness and consistency of POLS, RE, FE and FD: For you to remember: **consistency** of an estimator means that as the sample size gets large the estimate gets closer and closer to the true value of the parameter. **Unbiasedness** is a finite sample property that is not affected by increasing sample size. An estimate is unbiased if its expected value equals the true parameter value.
Let's put all the interesting assumptions that you might have encountered below and the let's try to attribute them to each model:

a. **Linearity**: $y_{it} = \beta_0 + \beta_1 x_{it1} + \beta_2 x_{it2} + \cdots + \beta_k x_{itk} + \alpha_i + u_{it} \longrightarrow$ we could have included also a time dependent term δ_t but we can get rid of it **by including time dummies in the model!**

b. Random sample in cross sections (no unobserved correlation among individuals);

c. $E[u_{it}|X_{is}, \alpha_i] = 0$. Now this is called **general exogeneity** assumption. This assumption is **strict** if we have the validity for present, past and future time, i.e. $E[u_{it}|X_{is}, \alpha_i] = 0$ for $s = 1, \ldots T$. Another weaker form of exogeneity is called **sequential** exogeneity which is valid for only present and future values of s. In other words, $E[u_{it}|X_{is}, \alpha_i] = 0$ for $s = t, t+1, \ldots T$.
The exogeneity assumption is central to decide between FE/FD and RE as we will see.

d. **No perfect collinearity**: there are no variables in X that are perfect linear combination one of the other. You might have seen this as requiring that the rank of X is full.

e. We must have **some time variation** in our explanatory variable X. In other words, X must be time dependent. Why? Otherwise, if it is only individual dependent FE would eliminate it!

f. $Cov(\alpha_i, X_{is}) = 0$

g. **Homoskedasticity**

h. **Non-autocorrelation** of the errors. In other words it is undesirable to have errors made like this: $u_{it} = \beta_1 u_{i,t-1} + \dots \beta_s u_{i,t-s} + \xi_{it}$. As we will see, FD can handle a specific type of non-autocorrelation, i.e. when $u_{it} = u_{i,t-1} + \xi_{it}$, i.e. the error is a random walk. For other cases we have to turn to panel GLS techniques (not covered).

Assumptions from a. to d. and h. are required by all panel model, i.e. POLS, FE, FD, RE. Assumption e. is require only for FD and FE. Assumption f. is required only for POLS and RE (as α_i appears in the error term). Assumption g. is required only for POLS.
Tab.(1) summarizes the above:

	POLS	**FE**	**FD**	**RE**
a.	Yes	Yes	Yes	Yes
b.	Yes	Yes	Yes	Yes
c.	Yes	Yes	Yes	Yes
d.	Yes	Yes	Yes	Yes
e.	No	Yes	Yes	No
f.	Yes	No	No	Yes
g.	Yes	No	No	No
h.	Yes	Yes	Yes*	Yes

* unless u_{it} is a random walk.

If the required assumptions are met then POLS will be consistent and unbiased. FE and FD will be consistent and unbiased **for** $N \longrightarrow \infty$ and **fixed** T. Notice that, RE instead is consistent but **not unbiased** because is in itself a feasible generalized least squares (we have to estimate $\hat{\lambda}$ rather than considering it as given).

Choosing between Pooled OLS and FE/RE: As we have seen above, there are six assumptions for simple **linear** regression models that must be fulfilled. Two of them can help us in choosing between Pooled OLS and FE/RE. Again, these assumptions are (1) Linearity, (2) Exogeneity, (3) Homoskedasticity and (3) Non-autocorrelation, (4) Independent variables are not Stochastic and (5) No Multicolinearity. If assumption (2) or (3) (or both) are violated, then FE or RE might be more suitable.

Choosing between FE and RE: Answering this question depends on your assumption, if the individual, unobserved heterogeneity is a constant or a random effect (i.e. it belongs to the error term). Specifically, if we see the model as

$$y_{it} = x_{it}\beta + (\alpha_i + u_{it})$$

, if α_i is correlated with x_{it} then RE (and POLS) are inconsistent and we have to use FE **provided that strict exogeneity holds.** What if strict exogeneity does not hold but sequential exogeneity does? Unfortunately neither of the

panel techniques (FE, RE, FD and POLS) can be used!
But this question can also be answered perfoming the Hausman-Test.
As you have seen in IV, the Hausman-Test is a test of endogeneity. By running the Hausman-Test, the null hypothesis is that the covariance between IV(s) and alpha is zero. If this is the case, then RE is preferred over FE. If the null hypothesis is not true, we must go with the FE-model.
In other words, in Eq.(1), if we assume for the moment that $\delta = 0$ and $\gamma = 1$, we are testing if the error $\nu_{it} = u_{it} + z_i$ correlates with X_{it} or not.

Pooled OLS (quick look)

Pooled OLS model **assumes** that the panel model looks like this:

$$Y_{it} = \beta X_{it} + uit, \quad i = 1, \ldots, N \quad t = \ldots, T$$

If u_{it} is uncorrelated with X_{it} (see assumptions above) we can estimate β consistently through OLS.
To do inference based on the conventional OLS estimator of the covariance matrix, we need to assume homoskedasticity and no serial correlation in the data. Both of these assumptions can be restrictive, especially the latter one. As a rule of thumb, it is a good idea to obtain an estimate of the covariance matrix that is robust to heteroskedasticity and autocorrelation, using the following sandwich formula:

$$V(\hat{\beta}^{POLS}) = \left(\sum_i X_i X_i' \right)^{(-1)} \left(\sum_i X_i \hat{u}_i \hat{u}_i' X_i' \right)^{(-1)} \left(\sum_i X_i X_i' \right)^{(-1)}$$

Random Effects (deep overview)

Let's re-write the model for convenience in this way:

$$y_{it} = x_{it}\beta + (\alpha_i + u_{it})$$

The very important assumptions for the model to work are the ones in the Table above. However, I will list the key ones below:

- α_i uncorrelated with x_{it}
- Strict exogeneity

Consider using POLS in this case. It is straightforward to show that POLS is inefficient since the residual $\nu^{POLS} = \alpha_i + u_i t$ is serially correlated:

$$\begin{aligned} &E[\nu_{it}^{POLS}, \nu_{it-s}^{POLS}] = E[(\alpha_i + u_i t)(\alpha_i + u_i t - s)] = \\ &E[\alpha_i^2 + \alpha_i u_{it} + \alpha_i u_{it-s} + u_{it} u_{it-s}] =^* \\ &E[\alpha_i^2] = \sigma_\alpha^2 \neq 0 \end{aligned}$$

* this follows from the assumption of non autocorrelation of u_{it}.
This implies that

$$corr(\nu_{it}^{POLS}, \nu_{it-s}^{POLS}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2}$$

If we are concerned with **efficiency**, we may want to consider a GLS estimator that takes this serial correlation into account. Also note that if σ_α^2 is high relative to σ_u^2 the serial correlation in the residual will be high. As a result the conventional estimator of the covariance matrix for the OLS estimator will not be correct.
RE is a GLS estimator solving the above problems! Using GLS involves transforming the original equation, so that the transformed equation fulfills the assumptions underlying the classical linear regression model. In other words **we want to manipulate the original equation so that it satisfies the OLS assumptions (in particular no serial correlation of errors).** We will transform the model and use OLS on the transformed model.
The panel data model is

$$y_{it} = x_{it}\beta + (\alpha_i + u_{it})$$

Define a

$$\lambda = 1 - \Big(\frac{\sigma_u^2}{T\sigma_\alpha^2 + \sigma_u^2}\Big)^{(1/2)}$$

Multiply λ by the individual average of the original equation:

$$y_{it} - \lambda\overline{y}_i = (x_{it} - \lambda\overline{x}_i)\beta + (\nu_{it}^{RE} - \lambda\overline{\nu}_i^{RE})$$

Using OLS on this, the transformed equation, gives the random effects GLS estimator. This estimator is efficient, because $(\nu_{it}^{RE} - \lambda\overline{\nu}_i^{RE})$ is serially uncorrelated (not proved here, just compute $E[(\nu_{it}^{RE} - \lambda\overline{\nu}_i^{RE}), (\nu_{it-s}^{RE} - \lambda\overline{\nu}_i^{RE})]$)!
The parameter λ is **unknown a priori**. This means that we have to estimaate its components, i.e. σ_u^2 and σ_α^2. There are various ways of doing this. The simplest, perhaps, is to use POLS in the first stage to obtain estimates of the composite residual $\hat{\nu}_{it}$. Based on this, we can calculate σ_α^2 as the covariance between $\hat{\nu}_{it}$ and $\hat{\nu}_{it-1}$ (see above the intuition). And, by definition:

$$\hat{\sigma}_u^2 = \hat{\sigma}_\nu^2 + \hat{\sigma}_u^2$$

Summarizing, to estimate λ we have a 2 steps procedure:

i. Estimate $\hat{\lambda}$ via POLS or FE. This means that you estimate the equation via, say, FE and you get $\hat{\alpha}$. after you got it, you compute $\hat{\sigma}_\alpha^2$.

ii. Use POLS on the transformed equation: $y_{it} - \hat{\lambda}\overline{y}_i = (x_{it} - \hat{\lambda}\overline{x}_i)\beta + (\nu_{it}^{RE} - \hat{\lambda}\overline{\nu}_i^{RE})$

This is usually done **automatically** in softwares.
Notice: if $\lambda = 1$ we obtain FE (this happens when the denominator of λ,

$\sigma_\alpha^2 \longrightarrow \infty$ meaning that α_i matters a lot and we have to eliminate it). If $\lambda = 0$ (this happens when $\sigma_\alpha^2 = 0$, i.e. α_i is unimportant and since u_{it} is uncorrelated with X_{it}, we can apply OLS) we obtain POLS.
Considerations: in RE we allow for time constant variables in X, but this means that we no longer get the nice property of eliminating the unobserved heterogeneity α_i.

Fixed Effects (few suggestions)

Assumptions:

1. α_i **freely** correlate with x_{it}
2. $E[x_{it}u_{is}] = 0 \quad s = 1, \ldots, T$

1. an 2. ensure **consistency** of FE. If 2. does not hold, we might turn to dynamic panels to find instruments for x_{it}.
Notice that FD requires for 2. a weaker exogeneity condition, i.e.: $E[x_{it}u_{is}] = 0 \quad s = t, t-1$ (e.g. we do not require that $E[x_{it}u_{i,t-2}] = 0$). This is because FD only subtracts $t-1$ values!

FE or FD? First of all, when $T = 2$ (i.e. we have only two time periods), FE and FD are exactly equivalent and so in this case it does not matter which one we use (try to prove this). But when $T \geq 3$, FE and FD are not the same. Under the null hypothesis that the model is correctly specified, FE and FD will differ only because of sampling error whenever $T \geq 3$. Hence, if FE and FD are significantly different - so that the differences in the estimates **cannot be attributed to sampling error** - we should worry about the validity of the strict exogeneity assumption.
If u_{it} is a random walk ($u_{it} = u_{i,t-1} + \xi_{it}$), then Δu_{it} is serially **uncorrelated** ($\Delta u_{it} = u_{it} - u_{i,t-1} = \xi_{it}$, where ξ_{it} is noise) and so the FD estimator will be more efficient than the FE estimator.
Conversely, under **"classical assumptions"** (this is why FE is usually more employed), i.e. $u_{it} \sim iid(0, \sigma_u^2)$, the FE estimator will be more efficient than the FD estimator (as in this case the FD residual Δu_{it} will exhibit negative **serial** correlation, since intuitively $\Delta u_{it} = u_{it} - u_{i,t-1}$ and $\Delta u_{it-1} = u_{it-1} - u_{i,t-2}$ so that each $u_{i,t-1}$ appears in the two with opposite sign making them negatively correlated. This did not matter when u_{it} was a random walk as we had $u_{it} - u_{i,t-1} = \xi_{it}$ and $u_{it-1} - u_{i,t-2} = \xi_{it-1}$, being ξ_{it} and ξ_{it-1} uncorrelated since they are pure noise!).

What if both α_i and δt appear in the model?

In this case we have to adopt time dummies.
Be sure to check for strict exogeneity to be present!
Furthermore think in terms of the **transformed** model.
I cannot say more on the topic (Professor told me so).

R command

```
fe <-plm(y~x, data=panel_df,model="within",index=c("id","tt"))

re <-plm(y~x, data=panel_df,model="random",index=c("id","tt"))
```

Notice that in order for the models to work, we need some time variability. Go to R exaample.

Clarifications on Logit and Probit

Unfortunately we have not enough time to cover them properly.
The key points are however summarized in the previous pdf that I send you. In case you need further clarifications, just send me an email.

References

[1] Wooldridge et al. (2010). Econometric analysis of Cross Section and Panel Data.