Introduction to Econometrics II: Recitation 7

Seung-hun Lee[*]

March 7th, 2022

1 Panel regression

1.1 Fixed effects regression

Let the data generating process at the individual-time observation level be[1]

$$y_{it} = x_{it}'\beta + c_i + e_{it}$$

Here, the unobserved fixed effect c_i is correlated with x_{it}. Our claim that OLS estimation method on POLS and RE methods being consistent relied on the fact that c_i is uncorrelated with x_{it}. With that broken down, we can no longer run an OLS on the above equation to get the consistent estimation for β. Therefore, it is essential that we minimize the role of c_i in the above equation in order to get the consistent estimation. We now introduce and apply new set of methods - fixed effects estimation - that achieves this goal.

There are three ways to approach fixed effects estimation - within estimation (WE), least square dummy variables (LSDV), and first difference (FD). We start with the within estimation method

1.1.1 Within estimation models

Within estimation attempts to weed out c_i by subtracting time averaged variables from the original data generating process. Write $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and similarly for other variables.

[*] Contact me at sl4436@columbia.edu if you spot any errors or have suggestions on improving this note.

[1] In Wooldridge (2010), the DGP has x_{it} and not x_{it}', x_{it} is a $1 \times k$ vector there, where it is $k \times 1$ vector here.

By averaging over time across all variables, we can get a cross-sectional equation which is transformed from the original data generating process

$$\bar{y}_i = \bar{x}_i'\beta + c_i + \bar{e}_i$$

The key is that even if y_i is average across time, we still get c_i itself. This is because $\frac{1}{T}\sum_{t=1}^{T} c_i = (Tc_i)/T = c_i$. So subtract the cross-sectional equation from the original data generating process to get

$$\ddot{y}_{it} = \ddot{x}_{it}'\beta + \ddot{e}_{it}, \ \ (i = 1,..,n, \text{ and } t = 1,..,T)$$

where $\ddot{y}_{it} = y_{it} - \bar{y}_i$. The within estimator is obtained by taking an OLS to above equation

$$\begin{aligned}\hat{\beta}_{WE} &= \left(\sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{x}_{it}\ddot{x}_{it}'\right)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{x}_{it}\ddot{y}_{it} \\ &= \left(\sum_{i=1}^{n}\ddot{X}_i'\ddot{X}_i\right)^{-1}\sum_{i=1}^{n}\ddot{X}_i'\ddot{y}_i\end{aligned}$$

Before we show further properties of the FE, we assume the following for the fixed effects estimators

Assumption 1.1 (Fixed effects assumptions)**.** *The following are the assumptions for the fixed effects models*

FE1 *We assume strict exogeneity* $E[e_{it}|X_i, c_i] = 0$

FE2 *rank* $\left(E[\ddot{X}_i'\ddot{X}_i]\right) = rank\left(\sum_{t=1}^{T} E[\ddot{x}_{it}'\ddot{x}_{it}]\right) = k$ *(full column rank)*

FE3 *Conditionally spherical variance matrix:* $E[e_ie_i'|X_i, c_i] = \sigma_e^2 I_T$

We first show that WE estimator is consistent. We rewrite the WE estimator os

$$\hat{\beta}_{WE} - \beta = \left(\sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{x}_{it}\ddot{x}_{it}'\right)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{x}_{it}\ddot{e}_{it}$$

$E[\ddot{x}_{it}\ddot{e}_{it}]$ can be written as

$$E[\ddot{x}_{it}\ddot{e}_{it}] = E[x_{it}e_{it}] - E[x_{it}\bar{e}_i] - E[\bar{x}_ie_{it}] + E[\bar{x}_i\bar{e}_i]$$

Since $\bar{x}_i, \bar{e}_i$ incorporates regressors and errors from all time periods, applying strict exogeneity (and strict exogeneity only) reduces the above equation to 0. Therefore, $\hat{\beta}_{WE}$ is consistent. Note that for this estimator, we REALLY need a strict exogeneity assumption. Anything weaker than this could make this estimator inconsistent. In addition, we need that $\sum_{t=1}^{T} E\left[\ddot{x}_{it}\ddot{x}_{it}'\right]$ be a full column rank for its inverse to be defined.

Assumption **FE3** implies that the variance-covariance of $\ddot{e}_{it}$ is computed as $(t \neq s)$

$$\begin{aligned} E[\ddot{e}_{it}^2] &= E[(e_{it} - \bar{e}_i)^2] = E[e_{it}^2] - 2E[e_{it}\bar{e}_i] + E[\bar{e}_i^2] \\ &= \sigma_e^2 - \frac{2}{T}\sigma_e^2 + \frac{1}{T}\sigma_e^2 = \sigma_e^2\left(1 - \frac{1}{T}\right) \\ E[\ddot{e}_{it}\ddot{e}_{is}] &= E[(e_{it} - \bar{e}_i)(e_{is} - \bar{e}_i)] = E[e_{it}e_{is}] - E[e_{it}\bar{e}_i] - E[e_{is}\bar{e}_i] + E[\bar{e}_i^2] \\ &= 0 - \frac{1}{T}\sigma_e^2 - \frac{1}{T}\sigma_e^2 + \frac{1}{T}\sigma_e^2 = -\frac{1}{T}\sigma_e^2 \end{aligned}$$

So $\ddot{e}_{it}$ is conditionally homoskedastic and have negative serial correlation (goes to 0 when T is large). Fortunately, it is known that due to the nature of time-demeaning, this causes only minor complications (Wooldridge 2010). The asymptotic distribution of the WE estimator is

$$\sqrt{n}(\hat{\beta}_{WE} - \beta) \sim N(0, E[\ddot{X}_i'\ddot{X}_i]^{-1}E[\ddot{X}_i'e_ie_i'\ddot{X}_i]E[\ddot{X}_i'\ddot{X}_i]^{-1})$$

where we can show $\ddot{X}_i'e_i = \ddot{X}_i'\ddot{e}_i$ using the definition of $\ddot{e}_i$ and strict exogeneity.

If we impose **FE3**, Then the asymptotic variance is $\sigma_e^2 E[\ddot{X}_i'\ddot{X}_i]^{-1}$. The estimator of the asymptotic variance would be $\widehat{\sigma}_e^2\left(n^{-1}\sum_{i=1}^{n}\ddot{X}_i'\ddot{X}_i\right)^{-1}$. To obtain $\widehat{\sigma}_e^2$, we start from our previous finding that $E[\ddot{e}_{it}^2] = \frac{T-1}{T}\sigma_e^2$. This implies that

$$\frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T}E[\ddot{e}_{it}^2] = \sigma_e^2$$

Then, we apply the (small-sample) correction by subtracting for k regressors. Thus, the estimate of σ_e^2 is

$$\widehat{\sigma}_e^2 = \frac{1}{n(T-1) - k}\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{\ddot{e}}_{it}$$

where $\widehat{\ddot{e}}_{it}$ is obtained from the OLS residual of the demeaned data generating process.

The within effects have a matrix-based notations as well. Stack up the individual-time level observation by each individuals to get

$$y_i = X_i\beta + 1_T c_i + e_i$$

where 1_T is the $T \times 1$ vector of $1'$s as elements. Now I define $Q_T \equiv I_T - 1_T(1_T' 1_T)^{-1} 1_T'$[2] Note that $(1_T' 1_T)^{-1} = T^{-1}$ and $1_T 1_T'$ is the T-dimensional square matrix of $1'$s as elements. Q_T is symmetric and idempotent. Now, premultiply Q_T to the indivudually-stacked data generating process to get

$$Q_T y_i = Q_T X_i \beta + Q_T 1_T c_i + Q_T e_i$$

A key feature is that $Q_T 1_T = 0$ and $Q_T y_i = \ddot{y}_i$. The latter is because

$$Q_T y_i = y_i - \frac{1}{T} \begin{pmatrix} 1 & 1 & ... & 1 \\ ... & ... & ... & ... \\ 1 & 1 & ... & 1 \end{pmatrix} \underbrace{\begin{pmatrix} y_{i1} \\ y_{i2} \\ .... \\ y_{iT} \end{pmatrix}}_{y_i} = y_i - \frac{1}{T} \begin{pmatrix} \sum_t y_{it} \\ ... \\ \sum_t y_{it} \end{pmatrix} = \begin{pmatrix} y_{i1} - \frac{1}{T}\sum_t y_{it} \\ ... \\ y_{iT} - \frac{1}{T}\sum_t y_{it} \end{pmatrix} = \ddot{y}_i$$

Thus, the transformed data generating process above is the demeaned data generating process. The WE estimator from this can be written as $\hat{\beta}_{WE} = \left(\sum_{i=1}^n X_i' Q_T X_i\right)^{-1} \sum_{i=1}^n X_i' Q_T y_i$.

1.1.2 Least-squares dummy variables

The idea behind LSDV models is that each c_i is a parameter to be estimated for each i - or that each i has distinct intercept. Let Dk_i be the dummy variable that equals 1 if $i = k$ and 0 otherwise. The idea is to put a total of $N - 1$ of such dummy variables into the regression[3]. Therefore, we work with

$$y_{it} = x_{it}'\beta + D1_i c_1 + ... + D(n-1)_i c_{n-1} + u_{it}$$

For the n'th individual, the constant term is represented by the β_0, the coefficient on the column vector of x_{it}. For $k(\neq n)$'th individual, the intercept term is $\beta_0 + c_k$.

Another way to characterize this is to further stack up y_i, X_i and e_i to get a nT-dimensional vector (or $nT \times k$ for X). Since c_i is different for each i, the idea here is to use a Kronecker product and express the individual fixed effect as

$$\begin{pmatrix} 1_T & ... & 0 \\ ... & ... & ... \\ 0 & ... & 1_T \end{pmatrix} \begin{pmatrix} c_1 \\ ... \\ c_n \end{pmatrix} = (I_n \otimes 1_T)c \in \mathbb{R}^{nT \times n} \times \mathbb{R}^{n \times 1} = \mathbb{R}^{nT \times 1}$$

[2] This is the residual matrix $M_X = I - P_X$.

[3] This is assuming that x_{it} contains a vector of 1's for a 'overall' constant.

Combine what we know to get

$$y = X\beta + \underbrace{(I_n \otimes 1_T)}_{=D}c + e$$

We then use the Frisch-Waugh-Lovell theorem to get $\hat{\beta}_{LSDV} = (X'M_D X)^{-1}(X'M_D y)$.

If we further use the properties of the Kronecker product, we can actually show that LSDV and WE are numerically identical.

Definition 1.1 (Kronecker Product)**.** Let $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{p\times q}$. Then their Kronecker product $A \otimes B$ is defined as

$$A \otimes B = \begin{pmatrix} a_{11}B & ... & a_{1n}B \\ ... & ... & ... \\ a_{m1}B & ... & a_{mn}B \end{pmatrix} \in \mathbb{R}^{mp\times nq}$$

where $B = \begin{pmatrix} b_{11} & ... & b_{1q} \\ ... & ... & ... \\ b_{p1} & ... & b_{pq} \end{pmatrix}$. Some of its properties are

- $A \otimes (B + C) = A \otimes B + A \otimes C$ (B, C are of same size)
- $(A \otimes B)(C \otimes D) = AC \otimes BD$ (Assuming multiplication is implementable)
- $(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1})$ (Assuming inverses are defined)
- $(A \otimes B)' = (A' \otimes B')$

Note that $M_D = I_{nT} - D(D'D)^{-1}D'$. Rewrite $D(D'D)^{-1}D'$ using the properties mentioned above as

$$\begin{aligned} D(D'D)^{-1}D' &= (I_n \otimes 1_T)[(I_n \otimes 1_T)'(I_n \otimes 1_T)]^{-1}(I_n \otimes 1_T)' \\ &= (I_n \otimes 1_T)[(I_n \otimes 1_T')(I_n \otimes 1_T)]^{-1}(I_n \otimes 1_T') \\ &= (I_n \otimes 1_T)[(I_n \otimes 1_T'1_T)]^{-1}(I_n \otimes 1_T') \\ &= (I_n \otimes 1_T)[I_n \otimes (1_T'1_T)^{-1}](I_n \otimes 1_T') \\ &= I_n \otimes 1_T(1_T'1_T)^{-1}1_T' \end{aligned}$$

Since I_{nT} can be written as $I_n \otimes I_T$, we can further write (use first property)

$$M_D = I_n \otimes (I_T - 1_T(1_T'1_T)^{-1}1_T') = I_n \otimes Q_T$$

So $\hat{\beta}_{LSDV}$ is equal to

$$\hat{\beta}_{LSDV} = (X'(I_n \otimes Q_T)X)^{-1}(X'(I_n \otimes Q_T)y)$$

Since $X = \begin{pmatrix} X_1 \\ ... \\ X_n \end{pmatrix}$, $I_n \otimes Q_T = \begin{pmatrix} Q_T & 0 & 0 \\ ... & ... & ... \\ 0 & 0 & Q_T \end{pmatrix}$ and $y = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix}$, we get

$$\hat{\beta}_{LSDV} = \left(\sum_{i=1}^{n} X_i' Q_T X_i \right)^{-1} \sum_{i=1}^{n} X_i' Q_T y_i = \hat{\beta}_{WE}$$

and we are done.

1.1.3 First-differenced models

The approach here is to take out c_i, a time-invariant unobserved fixed effect by taking differences of the observation over time. For this to work, we need $T \geq 2$. To obtain the first difference estimator, we need to subtract the original data generating process by one lag of y_{it}. Or

$$\Delta y_{it} = \Delta x_{it}' \beta + \Delta u_{it} \quad (i = 1, ..., n \text{ and } t = 2, .., T)$$

where $\Delta y_{it} = y_{i,t} - y_{i,t-1}$ and similarly for other variables. Notice that since c_i is same across $t = 1, .., T$ (but different for each i), it vanishes. By taking an OLS, we can obtain

$$\hat{\beta}_{FD} = \left(\sum_{i=1}^{n} \sum_{t=2}^{T} \Delta x_{it} \Delta x_{it}' \right)^{-1} \sum_{i=1}^{n} \sum_{t=2}^{T} \Delta x_{it} \Delta y_{it}$$

We set the following assumptions for the first differenced estimators

Assumption 1.2 (First differencing assumptions)**.** *The following are the assumptions for the first differenced models*

FD1 *We assume strict exogeneity* $E[e_{it}|X_i, c_i] = 0$

FD2 *rank* $\left(E[\Delta X_i' \Delta X_i]\right) = \textit{rank} \left(\sum_{t=2}^{T} E[\Delta x_{it} \Delta x_{it}']\right) = k$ *(full column rank)*

FD3 *Conditionally spherical variance matrix:* $E[\Delta e_i \Delta e_i' | X_i, c_i] = \sigma_{\Delta e}^2 I_{T-1}$

We can show that this is a consistent estimator. Write

$$\hat{\beta}_{FD} - \beta = \left(\sum_{i=1}^{n} \sum_{t=2}^{T} \Delta x_{it} \Delta x_{it}' \right)^{-1} \sum_{i=1}^{n} \sum_{t=2}^{T} \Delta x_{it} \Delta e_{it}$$

We need to show that $E[\Delta x_{it} \Delta e_{it}] = 0$. This can be written

$$\begin{aligned} E[\Delta x_{it} \Delta e_{it}] &= E[(x_{it} - x_{i,t-1})(e_{it} - e_{i,t-1})] \\ &= E[x_{it} e_{it}] - E[x_{it} e_{i,t-1}] - E[x_{i,t-1} e_{it}] + E[x_{i,t-1} e_{i,t-1}] \\ &= 0 - 0 - 0 + 0 = 0 \end{aligned}$$

Therefore, $\hat{\beta}_{FD}$ is consistent. Another requirement for this to be defined is that $\left(\sum_{t=2}^{T} \Delta x_{it} \Delta x_{it}' \right)$ should be a full column matrix so that the inverse matrix is defined.

So what is the difference between FD and WE estimators? One obvious difference is that with WE, we can still use T observation for each i where as we lose 1 period for each i in FD. The other relates to the structure of the error terms. If e_{it} is free from serial correlation (or e_t is an IID), then taking a FD would introduce serial correlation. This is because

$$\begin{aligned} cov(\Delta e_{it}, \Delta e_{i,t-1}) &= E[e_{it} e_{it-1}] - E[e_{it} e_{it-2}] - E[e_{it-1} e_{it-1}] + E[e_{it-1} e_{it-2}] \\ &= 0 - 0 - var(e_{it-1}) + 0 \neq 0 \end{aligned}$$

There may be a case when Δe_{it} is serially uncorrelated. For instance, e_{it} could be a random walk process in the sense that

$$e_{it} = e_{it-1} + \eta_{it} \;\; (E[\eta_{it}] = 0, E[\eta_{it} \eta_{is}] = 0 (s \neq t), var(e_{it}) = \sigma^2)$$

If we use a FD estimator here, we get to obtain the most efficient estimator by removing autocorrelation. WE, on the other hand, becomes inefficient because it does not get rid of the correlation structure in e_{it}.

In the case that $T = 2$, you can show that WE and FD estimators are numerically identical. Start with the WE, written as

$$\left(\sum_{i=1}^{n} \sum_{t=1}^{2} \ddot{x}_{it} \ddot{x}_{it}' \right)^{-1} \left(\sum_{i=1}^{n} \sum_{t=1}^{2} \ddot{x}_{it} \ddot{y}_{it} \right)$$

Each term can be replaced by

$$
\begin{aligned}
\sum_{t=1}^{2} \ddot{x}_{it}\ddot{x}_{it}' &= \sum_{t=1}^{2} \left(x_{it} - \frac{x_{i1}+x_{i2}}{2} \right) \left(x_{it} - \frac{x_{i1}+x_{i2}}{2} \right)' \\
&= \sum_{t=1}^{2} x_{it}x_{it}' - \frac{x_{i1}x_{i1}' + x_{i1}x_{i2}' + x_{i2}x_{i1}' + x_{i2}x_{i2}'}{2} \\
&= \frac{x_{i1}x_{i1}' - x_{i1}x_{i2}' - x_{i2}x_{i1}' + x_{i2}x_{i2}'}{2} \\
&= \frac{(x_{i1}-x_{i2})(x_{i1}-x_{i2})'}{2} = \frac{\Delta x_{i2}\Delta x_{i2}'}{2}
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_{t=1}^{2} \ddot{x}_{it}\ddot{y}_{it} &= \sum_{t=1}^{2} \left(x_{it} - \frac{x_{i1}+x_{i2}}{2} \right) \left(y_{it} - \frac{y_{i1}+y_{i2}}{2} \right) \\
&= \sum_{t=1}^{2} x_{i}y_{i} - \frac{x_{i1}y_{i1} + x_{i1}y_{i2} + x_{i2}y_{i1} + x_{i2}y_{i2}}{2} \\
&= \frac{x_{i1}y_{i1} - x_{i1}y_{i2} - x_{i2}y_{i1} + x_{i2}y_{i2}}{2} \\
&= \frac{(x_{i1}-x_{i2})(y_{i1}-y_{i2})}{2} = \frac{\Delta x_{i2}\Delta y_{i2}}{2}
\end{aligned}
$$

Combine the two to get $\left(\sum_{i=1}^{n} \Delta x_{i2}\Delta x_{i2}'\right)^{-1} \sum_{i=1}^{n} \Delta x_{i2}\Delta y_{i2}$, the FD estimator when $T = 2$.

1.2 Time-invariant regressors in the fixed effects regression

For all fixed effects methods, the common pitfall is that it is impossible to include time-invariant regressors. This is because they are either erased due to the transformation process (WE, FD) or absorbed by a separate variable for unobserved individual effects (LSDV). While they are generally not a problem, it becomes a major one when our variable of interest is time-invariant: gender and race, for instance. In this situation, Hausman and Taylor (1981) propose an estimation approach based on method of moments that can identify the effects of the time-invariant variables.

Write the data generating process as

$$
y_{it} = z_i'\gamma + x_{it}'\beta + c_i + e_{it}
$$

where we are interested in γ. Assume $E[z_i e_{it}] = 0, E[x_{it} e_i] = 0, E[z_i c_i] = 0$, and $E[x_{it} c_i] \neq 0$

(or strict exogeneity for e_{it}). Then, we have two moment conditions that we can work with

$$E[\ddot{x}_{it}e_{it}] = E[\ddot{x}_{it}(y_{it} - x'_{it}\beta - z'_i\gamma)] = 0$$
$$E[z_i e_{it}] = E[z_i(y_{it} - x'_{it}\beta - z'_i\gamma)] = 0$$

Thus, the valid IV for this procedure is $\ddot{x}_{it}$ and z_i. We can obtain β and γ estimates by a 2SLS procedure with $\ddot{x}_{it}$ and z_i as the set of IVs.

There is another algebraically equal process, according to Hausman and Taylor (1981). The steps are as follows:

1. Use within estimation to obtain $\hat{\beta}$
2. Obtain $\hat{\gamma} = (Z'Z)^{-1}Z'\hat{c}$ (where Z is z_{it} for the full sample level and). From the moment condition on z_i, we can get

   $$z'_i(y_{it} - x'_{it}\beta - z_i\gamma) = 0$$

   Replace β with $\hat{\beta}$ from the previous step. Then note that

   $$z'_i(y_{it} - x'_{it}\hat{\beta} - z_i\gamma) = z'_i(\bar{y}_i - \bar{x}'_i\hat{\beta} - z_i\gamma) = 0$$

   where $z'_i(y_{it} - x'_{it}\hat{\beta} - z_i\gamma) = 0$ is from the moment condition, $z'_i(\bar{y}_i - \bar{x}'_i\hat{\beta} - z_i\gamma) = 0$ is from taking averages across time on $z'_i(y_{it} - x'_{it}\hat{\beta} - z_i\gamma) = 0$. $\bar{e}_i$ can be ruled out since the moment condition $E[z_i e_{it}] = 0$ means we can erase $\bar{e}_i$ here. Combining the information we have, we use method of moments to get

   $$\frac{1}{n}\sum_{i=1}^{n} z'_i(\bar{y}_i - \bar{x}'_i\hat{\beta} - z_i\gamma) = 0 \iff \frac{1}{n}\sum_{i=1}^{n} z_i(\bar{y}_i - \bar{x}'_i\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} z_i z'_i\gamma$$

   The estimator for γ can thus be obtained by as $\hat{\gamma} = \left(\sum_{i=1}^{n} z_i z'_i\right)^{-1}\left(\sum_{i=1}^{n} z_i(\bar{y}_i - \bar{x}_i\hat{\beta})\right)$. In practice, this is obtained by regressing $\hat{c}_i$ on z_i and thus $\hat{\gamma} = (Z'Z)^{-1}Z'\hat{c}$

Comment 1.1 (Between effects). *The expression $\bar{y}_i = \bar{x}'_i\beta + c_i + \bar{e}_i$ is the grouped mean (over time) for the DGP $y_{it} = x'_{it}\beta + c_i + e_{it}$. From the group mean, we can get the between estimator that uses variation in the mean of each group i around the overall mean.*

$$\hat{\beta}_{BE} = \left(\sum_{i=1}^{n} \bar{x}_i\bar{x}'_i\right)^{-1}\left(\sum_{i=1}^{n} \bar{x}_i\bar{y}_i\right)$$

However, this is rarely in use because the within group variation is a lost information in this process, leading to inefficient estimates. Moreover, c_i is not obliterated in the transformed DGP. So if c_i is correlated with x_{it}, it is then naturally correlated with $\bar{x}_i$ and between effects estimate is inconsistent.

1.3 Selecting between FE and RE

To determine the use of random effects vs fixed effects for estimating β, we note the following properties of FE and RE

- FE: always consistent, but inefficient if $E[X_i' c_i] = 0$
- RE: consistent and efficient if $E[X_i' c_i] = 0$, but otherwise inconsistent.

Because of these properties, we can apply a Hausman principle to test the use of random and fixed effects. We create this test statistic for the null hypothesis of $H_0 : E[X_i' c_i] = 0$

$$H \equiv (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[\widehat{V}_{\beta_{FE}-\beta_{RE}}]^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \sim \chi^2_{\dim(X)}$$

where we can write $\widehat{V}_{\beta_{FE}-\beta_{RE}} = \widehat{V}_{\beta_{FE}} - \widehat{V}_{\beta_{RE}}$. If the null is not rejected, then using RE is acceptable. Otherwise, RE is inconsistent and FE should be preferred.

1.4 Generalizing the structure of fixed effects

We have assume a one-way fixed effect structure where our only unobserved fixed effect is c_i. We can further generalize to include more fixed effects. For instance, we can include the unobserved time fixed effects δ_t that may also be correlated with x_{it} by

$$y_{it} = x_{it}'\beta + c_i + \delta_t + e_{it}$$

The fixed effects estimator should get rid of both c_i and δ_t. This is achievable with a two-step demeaning procedure. Define

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it},\ \bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_{it},\ \bar{y} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it}$$

Then the demeaning process used here is $\tilde{y}_{it} = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}$. That is because

$$\begin{aligned}\tilde{y}_{it} &= (\alpha_i + \gamma_t + x_{it}\beta + e_{it}) - (\alpha_i + \bar{\gamma} + \bar{x}_i\beta + \bar{e}_i) - (\bar{\alpha} + \gamma_t + \bar{x}_t\beta + \bar{e}_t) + (\bar{\alpha} + \bar{\gamma} + \bar{\bar{x}}\beta + \bar{\bar{e}}) \\ &= (x_{it} - \bar{x}_i - \bar{x}_t + \bar{\bar{x}})\beta + (e_{it} - \bar{e}_i - \bar{e}_t + \bar{\bar{e}}) \\ &= \tilde{x}_{it}\beta + \tilde{e}_{it}\end{aligned}$$

Then, the pooled OLS on the above equation would lead to consistent estimates of β.

We can generalize further. We can model fixed effects flexibly using interactive fixed effects where our composite error v_{it} can be written as $v_{it} = \lambda_i f_t + e_t$. f_t is a vector of factors, and λ_i is a vector of factor loadings. In fact, we can model two way fixed effects by letting $\lambda_i = \begin{pmatrix} 1 \\ c_i \end{pmatrix}$ and $f_t = \begin{pmatrix} \delta_t \\ 1 \end{pmatrix}$. With interactive fixed effects, we can model for unobservable individual effects that may vary over time by putting a fixed effect for each individual-time level at the cost of increasing computational burden.

1.5 Incidental parameters problem

In their 1948 Econometrica paper, Neyman and Scott introduced puzzling examples of MLE delivering an inconsistent estimates of the structural parameters[4]. The simplest formulation is

$$y_{it} = c_i + e_{it}, \quad e_{it} \sim iid\ N(0, \sigma_e^2)$$

The MLE estimate for the σ_e^2 is

$$\frac{1}{nT}\sum_i \sum_t (y_{it} - \bar{y}_i)^2 \sim \sigma^2 \chi^2_{n(T-1)}/nT$$

However, as we have shown for the fixed effects $E[\hat{\sigma}_e^2] = E[\ddot{e}_{it}^2] = \frac{T-1}{T}\sigma_e^2$. This means that our MLE estimate for σ_e^2 is inconsistent given a fixed T (Note that our asymptotics were based on $n \to \infty$). The key takeaway is that for a case where our variation in T is very limited, in that there are few time periods in the data, then we should be cautious about the consistency of our estimates.

[4] This part is from the lecture notes of Roger Koenker of UIUC, the same person who devised the quantile regressions. You can find the notes here (link).

2 Dynamic panel data

There are many instances where we are interested in how the past outcome affects the present and future outcomes - whether it be growth rate or factor demands. The data generating processes look like this

$$y_{it} = y_i^{t-1}\rho + x_i^t\beta + c_i + e_{it}$$

where y_i^{t-1} is a vector of $y_{i1}, ..., y_{it-1}$, x_i^t is a vector of $x_{i1}, ..., x_{it}$. For simplicity, we will focus on AR(1) models, so no x_i's and include only y_{it-1}. We also assume that $c_i \sim N(0, \sigma_c^2), e_{it} \sim N(0, \sigma_e^2)$ and that both are independent of each other. Although these are the similar assumption we set for fixed effects, the methods used there cannot estimate ρ consistently.

For POLS, where $v_{it} = c_i + e_{it}$, The OLS estimates would be

$$\hat{\rho} = \left(\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it-1}^2\right)^{-1} \sum_{i=1}^{n}\sum_{t=1}^{T} y_{it-1}y_{it}$$

or

$$\hat{\rho} - \rho = \left(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it-1}^2\right)^{-1} \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it-1}(c_i + e_{it})$$

Since $y_{it-1} = \rho y_{it-2} + c_i + e_{it-1}$, the $\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=0}^{T} y_{it-1}c_i$ term does not converge in probability to 0, leading to inconsistencies.

Even if we get rid of c_i by using the first difference at $T = 2$, ρ estimates are still inconsistent. To show this, start with

$$\Delta y_{i2} = \rho\Delta y_{i1} + \Delta e_{i2}$$

We can now show that the regressor and the error term are correlated, since

$$\begin{aligned} cov(\Delta y_{i1}, \Delta e_{i2}) &= E[(y_{i1} - y_{i0})(e_{i2} - e_{i1})] \\ &= E[y_{i1}e_{i2}] - E[y_{i1}e_{i1}] - E[y_{i0}e_{i2}] + E[y_{i0}e_{i1}] \end{aligned}$$

Because $E[y_{i1}e_{i1}] = E[(\alpha y_{i0} + c_i + e_{i0})e_{i0}]$, this contains term that is nonzero. So Δy_{i1} is an endogenous regressor.

Even for the within estimator, which is written as

$$y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it} = \rho\left(y_{it-1} - \frac{1}{T}\sum_{t=1}^{T} y_{it-1}\right) + e_{it} - \frac{1}{T}\sum_{t=1}^{T} e_{it}$$

The regressor contains $y_{i0}, .., y_{iT-1}$ and residuals contain $e_{i1}, .., e_{iT}$. There are overlapping time periods, implying that the regressor becomes endogenous.

2.1 Assumptions for the dynamic panels

To analyze dynamic panel model estimators, we require the following assumptions

Assumption 2.1 (Dynamic panel assumptions)**.** *For the DGP*

$$y_{it} = \rho y_{it-1} + x_{it}'\beta + c_i + e_{it}$$

where we use $w_{it} = (y_{it-1}\ x_{it}')$, the following are the assumptions for the dynamic panel models.

DP1 *Sequential exogeneity: $E[e_{it}|w_{it}, ..., w_{i1}, c_i] = 0$ for each t*

- *Or $E[w_{is}e_{it}] = 0$ for $s \in \{1, .., t\},\ t \in \{1, .., T\}$*

DP2 *Dynamic completeness: $E[y_{it}|x_t, y_{it-1}, x_{it-1}, y_{it-2}, ..., c_i] = E[y_{it}|x_t, y_{it-1}, c_i]$. This implies that x_t, y_{it-1} are all the lags needed and no information is lost by not including further lags. This implies that $E[e_{it}|x_t, y_{it-1}, x_{it-1}, y_{it-2}, ..., c_i] = 0$, or no residual correlation*

If w_{it} does not include lagged dependent variables, then dynamic completeness implies strict exogeneity. If w_{it} has lagged dependent variables, both conditions are the same condition (Wooldridge 2016). Since our case is the latter, we can use the term interchangeably.

The key difference is that we need to rely on sequential exogeneity, not strict exogeneity. In fact, it is difficult to take a stance on strict exogeneity. The inclusion of the lagged variable introduces feedbacks in which x_{it} can be affected by past values of y_{it}, say y_{it-1}. In such case, the past shock e_{it-1} can affect values of x_{it}. Thus a flexible exogeneity assumption that takes into account these feedback effects are needed.

The sequential exogeneity assumption implies the following

- For $s \leq t$, $E[w_{is}'e_{it}] = 0$
- For $s < t$, $E[e_{is}e_{it}] = 0$

2.2 Internal instrument approach

The sequential exogeneity assumption allows us to generate various moment conditions that do not require instrumental variables from outside the model. For example, we can have $E\left[\begin{pmatrix} x_{i1} \\ x_{i2} \\ y_{i1} \end{pmatrix} \Delta e_{i3}\right] = 0$ based on the condition that $E[e_{it}|x_{it}, y_{it-1}, x_{it-1}, y_{it-2}, ..., c_i] = 0$. So x_{i1}, x_{i2} and y_{i1} are candidates for a valid IV. Depending on the combination of the candidates selected, there are many potential estimators for the dynamic panel data.

This discussion is not limited to panel data. Go back to the non-panel AR(1) for the moment where

$$y_t = \alpha y_{t-1} + e_t, \ \ e_t \sim MA(1) = u_t + \theta u_{t-1}, \ u_t \sim WN$$

In such case, $E[y_{t-1}e_t]$ is nonzero. This is because

$$\begin{aligned} E[y_{t-1}e_t] &= E[(\alpha y_{t-2} + e_{t-1})e_t] \\ &= E[e_{t-1}e_t] \\ &= E[(u_{t-1} + \theta u_{t-2})(u_t + \theta u_{t-1})] = \theta var(u_{t-1}) \end{aligned}$$

But if we look at $E[y_{t-2}e_t]$,

$$\begin{aligned} E[y_{t-2}e_t] &= E[(\alpha y_{t-3} + e_{t-2})e_t] \\ &= E[e_{t-2}e_t] \\ &= E[(u_{t-2} + \theta u_{t-3})(u_t + \theta u_{t-1})] = 0 \end{aligned}$$

So y_{t-2} can be used to instrument y_{t-1}.

2.3 Anderson-Hsiao estimator

In estimating

$$y_{it} = \rho y_{it-1} + \alpha_i + e_{it}$$

Anderson and Hsiao (1982) proposed the following IV approach. First difference the above equation and obtain

$$\Delta y_{it} = \rho \Delta y_{it-1} + \Delta e_{it}$$

where possible values of i remain the same but $t = 2, ..., T$. The suggested IV is that Δy_{it-1} be instrumented with y_{it-2}. We can check that this is a valid IV.

- **Relevancy**: Note that $\Delta y_{it-1} = y_{it-1} - y_{it-2}$. So this term contains y_{it-2} and thus relevancy is satisfied.

- **Exogeneity**: Note that

$$\begin{aligned} cov(y_{it-2}, \Delta e_{it}) &= E[y_{it-2}, e_{it} - e_{it-1}] \\ &= E[y_{it-2}e_{it}] - E[y_{it-2}e_{it-1}] \\ &= 0 - 0 = 0 \end{aligned}$$

  The last line is justified as follows. Assuming $E[y_{i0}e_{it}] = 0$ for $t \geq 1$, we can expand to (or that y_{i0} is given)

$$\begin{aligned} E[y_{i1}e_{it}]\ (t \geq 2) &= E[(\rho y_{i0} + \alpha_i + e_{i1})e_{it}] \\ &= \rho E[y_{i0}e_{it}] + E[\alpha_i e_{it}] + E[e_{i1}e_{it}] \\ &= \rho \times 0 + 0 + 0 = 0 \end{aligned}$$

  Likewise,

$$\begin{aligned} E[y_{i2}e_{it}]\ (t \geq 3) &= E[(\rho y_{i1} + \alpha_i + e_{i2})e_{it}] \\ &= \rho E[y_{i1}e_{it}] + E[\alpha_i e_{it}] + E[e_{i2}e_{it}] = 0 \end{aligned}$$

  Therefore, we can generalize to $E[y_{is}e_{it}] = 0$ for $t > s$.

We can generalize using matrix notation. Define

$$\Delta y_i = \rho \Delta y_{i,-1} + \Delta e_i$$

where $\Delta y_i = \begin{pmatrix} \Delta y_{i2} \\ ... \\ \Delta y_{iT} \end{pmatrix}$, $\Delta y_{i,-1} = \begin{pmatrix} \Delta y_{i1} \\ ... \\ \Delta y_{iT-1} \end{pmatrix}$ and $\Delta e_i = \begin{pmatrix} \Delta e_{i2} \\ ... \\ \Delta e_{iT} \end{pmatrix}$. The matrix of instruments Z_i would be

$$Z_i = \begin{pmatrix} y_{i0} & 0 & 0 & ... \\ 0 & y_{i1} & 0 & ... \\ ... & ... & ... & ... \\ 0 & 0 & ... & y_{iT-2} \end{pmatrix} \in \mathbb{R}^{(T-1)\times(T-1)}$$

This is a just identified case in the sense that the number of IV is equal to the number of endogenous variables. So use the method of moments approach and solve

$$\begin{aligned} E[Z_i'\Delta e_i] = 0 &\iff E[Z_i'(\Delta y_i - \rho \Delta y_{i,-1})] = 0 \\ &\implies \frac{1}{n}\sum_{i=1}^{n}(Z_i'\Delta y_i - \rho Z_i'\Delta y_{i,-1}) = 0 \\ &\iff \rho\frac{1}{n}\sum_{i=1}^{n} Z_i'\Delta y_{i,-1} = \frac{1}{n}\sum_{i=1}^{n}\rho Z_i'\Delta y_i \\ &\iff \hat{\rho} = \left(\sum_{i=1}^{n} Z_i'\Delta y_{i,-1}\right)^{-1}\sum_{i=1}^{n} Z_i'\Delta y_i \end{aligned}$$

2.4 Arellano-Bond estimator

This approach is similar to that of Anderson-Hsiao in the sense that we start with the first differentiation. The difference is in the instruments used. Arellano and Bond (1991) suggests that to instrument for Δy_{it-1}, we use $y_{i0}, ..., y_{it-2}$ as instruments. To see why they are valid

- **Relevancy:** It should be clear why y_{it-2} is relevant. As for others, since $y_{it-1} = \rho y_{it-2} + u_{it-1}$ and $y_{it-2} = \rho y_{it-3} + u_{it-2}$, we can write recursively that

  $$y_{it-1} = \rho^2 y_{it-3} + \rho e_{it-2} + e_{it-1}$$

  ... and so on. Therefore, we can verify relevancy.

- **Exogeneity:** Note that $cov(y_{is}, \Delta e_{it})$ for any $s < t-1$ is 0, as we have shown above. So exogeneity holds as well.

The generalized approach using matrix will be similar except for the instrument matrix Z_i.

$$Z_i = \begin{pmatrix} y_{i0} & 0 & 0 & ... \\ 0 & (y_{i0}, y_{i1}) & 0 & ... \\ ... & ... & ... & ... \\ 0 & 0 & ... & (y_{i0}, ..., y_{iT-2}) \end{pmatrix} \in \mathbb{R}^{(T-1)\times\frac{T(T-1)}{2}}$$

Unlike Anderson-Hsiao estimator, this is an overidentified case. So we would need to use a GMM criterion with a weight matrix W_n. This would result in

$$\hat{\rho} = \arg\min_{\rho} \left\{ n \times \frac{1}{n}\sum_{i=1}^{n}(Z_i'\Delta y_i - \rho Z_i'\Delta y_{i,-1})' W_n \frac{1}{n}\sum_{i=1}^{n}(Z_i'\Delta y_i - \rho Z_i'\Delta y_{i,-1}) \right\}$$

$$\implies \left[\left(\sum_{i=1}^{n} \Delta y_{i,-1}' Z_i \right) W_n \left(\sum_{i=1}^{n} Z_i' \Delta y_{i,-1} \right) \right]^{-1} \left(\sum_{i=1}^{n} \Delta y_{i,-1}' Z_i \right) W_n \left(\sum_{i=1}^{n} Z_i' \Delta y_i \right)$$

If we have x_{it} variables in the model, we can also include those in the Z_i matrix. This depends on the assumption we set on $E[e_{it}|X_i]$'s.

The remaining question is to now select a matrix W_n that would lead to the lowest variance possible. If $g(Z_i, \rho)$ is the moment condition, the following would qualify as the most efficient weighting matrix.

$$W_n = E[g(Z_i, \rho) g(Z_i, \rho)']^{-1}$$

In our context, $g(Z_i, \rho)$ would be equivalent to $Z_i'\Delta e_i$. So we need a sample analogue of

$$E[Z_i'\Delta e_i \Delta e_i' Z_i] \implies \frac{1}{n}\sum_{i=1}^{n} Z_i \Delta e_i \Delta e_i' Z_i$$

For further simplification, we can assume two types of settings

- e_{it} is **IID and homoskedastic**: In such case, where $E[e_{it}^2] = \sigma_u^2$, we can write

$$E[\Delta e_i \Delta e_i] = E \begin{bmatrix} \Delta e_{i2}^2 & \Delta e_{i2}\Delta e_{i3} & ... & \Delta e_{i2}\Delta e_{iT} \\ \Delta e_{i2}\Delta e_{i3} & \Delta e_{i3}^2 & ... & \Delta e_{ie}\Delta e_{iT} \\ ... & ... & ... & ... \\ \Delta e_{i2}\Delta e_{iT} & \Delta e_{i3}\Delta e_{iT} & ... & \Delta e_{iT}^2 \end{bmatrix} = \begin{bmatrix} 2\sigma_u^2 & -\sigma_u^2 & ... & 0 \\ -\sigma_u^2 & 2\sigma_u^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 2\sigma_u^2 \end{bmatrix}$$

Define matrix $H \in \mathbb{R}^{(T-1)\times(T-1)}$ to have 2's in the diagonal elements, -1's in the immediate off-diagonals, and 0 everywhere else. Then $E[\Delta e_i \Delta e_i'] = \sigma_u^2 H$. This implies that the weighting matrix that we are looking for is

$$E[Z_i'\Delta e_i \Delta e_i' Z_i]^{-1} = \left(\frac{1}{n}\sum_{i=1}^{n} Z_i H Z_i \right)^{-1}$$

where σ_u^2 is taken out since scaling W_n by a scalar does not affect the value of the estimator. However, note that it may affect the test statistics for the overidentification test we will conduct later.

- e_{it} is **heteroskedastic**: For this, we take an approach similar to the two-step GMM estimator we did some weeks ago. The optimal weighting matrix in this case would be

$$W_n = \left(\frac{1}{n} \sum_{i=1}^{n} Z_i \Delta \tilde{e}_i \Delta \tilde{e}_i' Z_i \right)^{-1}$$

  where $\Delta \tilde{e}_i = \Delta y_i - \tilde{\rho} \Delta y_{i,-1}$, a residual from the preliminary estimator $\tilde{\rho}$. The preliminary estimator could be either from $W_n = I_{T(T-1)/2}$ or from the one-step GMM estimator that we derived earlier.

One thing to note is that because we are using more moment conditions that the number of endogenous variables, this is when we could test for an overidentification restriction. Suppose that W_n is the efficient weighting matrix. Then, similar to the GMM overidentification test, we are testing

$$H_0 : E[g(Z_i, \rho)] = 0, \quad H_1 : E[g(Z_i, \rho)] \neq 0$$

We can construct the following test statistic (Sargan-Hansen overidentification statistic)

$$J = n \bar{g}_n(\hat{\rho})' W_n \bar{g}_n(\hat{\rho})$$

Under H_0, J has a limiting distribution $\chi^2_{\left(\frac{T(T-1)}{2} - 1\right)}$. We lose one degree of freedom since we have used one estimator for ρ.