Introduction to Econometrics II: Recitation 5[*]

Seung-hun Lee[†]

February 19th, 2020

1 Topics on GMM

1.1 Wald Test Statistics

Assume that you have found any kind of a GMM estimator. So we are not limiting our discussion to the efficient GMM. The GMM estimator $\hat{\beta}_{GMM}$ that you found has a limiting distribution that can be characterized as

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) \xrightarrow{d} N(0, V_\beta)$$

Now, we define a fuction $r(\beta) : \mathbb{R}^k \xrightarrow{p} \Theta \in \mathbb{R}^q$ that characterizes the type of limitations we put on our parameter of interest β. I will write $\theta = r(\beta)$. The GMM estimator of θ would naturally be $\hat{\theta}_{GMM} = r(\hat{\beta}_{GMM})$. Then, by delta method, we can characterize the limiting distribution of $\hat{\theta}_{GMM}$ as

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta) \xrightarrow{d} R'N(0, V_\beta) = N(0, \underbrace{R'V_\beta R}_{=V_\theta})$$

where $R = \frac{\partial r(\beta)'}{\partial \beta} \in \mathbb{R}^{k \times q}$.

The hypothesis test can be set up in the following manner

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0$$

[*] This is based on the lecture notes of Professors Jushan Bai and Bernard Salanie. I was also greatly helped by previous recitation notes from Paul Sungwook Koh and Dong Woo Hahm. The remaining errors are mine.

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

Here, we make use of the Wald test statistic, defined as

$$W \equiv n(\hat{\theta} - \theta)' \widehat{V}_{\theta}^{-1} (\hat{\theta} - \theta)$$

where $W \xrightarrow{d} \chi^2_q$ under H_0. If we conduct a test with a significance level α, we need to find a critical value C such that $\alpha = 1 - F(C)$ where F is the CDF for χ^2_q. In such test, we reject H_0 if $W > C$ and do not reject if otherwise.

1.2 Restricted GMM

There are many cases where we may want to impose restrictions on the coefficients (e.g. hypothesis testing). In this section, we will consider $r(\beta) = 0$ as a constraint on β. Finding a restricted GMM (or constrained GMM) is not too different from the unconstrained GMM we have been solving so far. The only difference is that we are now solving a constrained minimization problem. The constrained GMM estimator, in math, is the solution to the following problem

$$\hat{\beta}_{CGMM} = \arg\min_{\beta} J_n(\beta) \text{ s.t. } r(\beta) = 0$$

For instance, we may consider a linear constraint on β coefficients in the form of $R'\beta = c$, where $R \in \mathbb{R}^{k \times q}$ with q being number of constraints. We can write this out in a Lagrangian form.

$$n\bar{g}(\beta)' W \bar{g}(\beta) + \lambda'[R'\beta - c]$$

where $\lambda \in \mathbb{R}^q$ is the vector of Lagrange multipiers.

Suppose that the moment condition given is $E(z_i e_i) = 0$. Then

$$\bar{g}(\beta) = \left(\frac{Z'y}{n} - \frac{Z'X\beta}{n} \right)$$

To find the optimal β, we differentiate the objective function with respect to β. This will become

$$\frac{y'ZWz'Y}{n} - \frac{y'ZWZX\beta}{n} - \frac{\beta'X'ZWZ'y}{n} + \frac{\beta'X'ZWZ'X\beta}{n} + \lambda'R'\beta - \lambda'C$$
$$\xrightarrow{\partial\beta} -2\frac{X'ZWZ'y}{n} + 2\frac{X'ZWZ'X\beta}{n} + R\lambda = 0$$

We can use the knowledge that the unconstrained GMM can be written as $\left(\frac{X'ZWZ'X}{n}\right)^{-1} \frac{X'ZWZ'y}{n}$.

So the trick is to pre-multiply $-R'\left(\frac{X'ZWZ'X}{n}\right)^{-1}$. This would change the first order condition to

$$2R'\hat{\beta}_{GMM} - 2R'\beta - R'\left(\frac{X'ZWZ'X}{n}\right)^{-1} R\lambda = 0$$

Since $R'\beta = c$ under H_0, we write

$$\lambda = 2\left(R'\left(\frac{X'ZWZ'X}{n}\right)^{-1} R\right)^{-1} R'(\hat{\beta}_{GMM} - \beta)$$

We can then put this back into the λ in the first order condition to get

$$-2\frac{X'ZWZ'y}{n} + 2\frac{X'ZWZ'X\beta}{n} + 2R\left(R'\left(\frac{X'ZWZ'X}{n}\right)^{-1} R\right)^{-1} R'(\hat{\beta}_{GMM} - \beta) = 0$$

Note that in this equation, n is commonly divided in all terms and 2 is multiplied in all terms. So we can practically get rid of them. So we can further rewrite the above as

$$(X'ZWZ'X)\beta - R(R'(X'ZWZ'X)^{-1}R)^{-1}\underbrace{R'\beta}_{=c} = X'ZWZ'y - R(R'(X'ZWZ'X)^{-1}R)^{-1}R'\hat{\beta}_{GMM}$$

Or

$$\begin{aligned}(X'ZWZ'X)\beta &= X'ZWZ'y - R(R'(X'ZWZ'X)^{-1}R)^{-1}(R'\hat{\beta}_{GMM} - c)\\ \Longrightarrow\ \hat{\beta}_{CGMM} &= \hat{\beta}_{GMM} - (X'ZWZ'X)^{-1}R(R'(X'ZWZ'X)^{-1}R)^{-1}(R'\hat{\beta}_{GMM} - c)\end{aligned}$$

Note that if we premultiply R' to both sides, we can get

$$\begin{aligned}R'\hat{\beta}_{CGMM} &= R'\hat{\beta}_{GMM} - R'(X'ZWZ'X)^{-1}R(R'(X'ZWZ'X)^{-1}R)^{-1}(R'\hat{\beta}_{GMM} - c)\\ &= R'\hat{\beta}_{GMM} - R'\hat{\beta}_{GMM} + c = c\end{aligned}$$

which shows that the CGMM satisfies the constraint.

1.3 Distance Test

When we looked at the Wald Test Statistics, we assumed a linear form of constraints for $r(\beta) = \theta$. If we are working with a possibly nonlinear form of $r(\beta)$, we can use an alternative criterion-based statistic. This is where the GMM Distance statistic comes in. The basic idea is to compare unrestricted and restricted estimators by contrasting the criterion functions. Again, we define

$$J(\beta) = n\bar{g}_n(\beta)'\widehat{\Omega}^{-1}\bar{g}_n(\beta)$$

where $\bar{g}_n(\beta)$ is $\frac{1}{n}\sum_{i=1}^{n} g(w_i, \beta)$ and $\widehat{\Omega}$ is the efficient weight matrix. With the unconstrained estimator and the constrained estimator, we can write

$$\begin{aligned} J(\hat{\beta}_{GMM}) &= n\bar{g}_n(\hat{\beta}_{GMM})'\widehat{\Omega}^{-1}\bar{g}_n(\hat{\beta}_{GMM}) \\ J(\hat{\beta}_{CGMM}) &= n\bar{g}_n(\hat{\beta}_{CGMM})'\widehat{\Omega}^{-1}\bar{g}_n(\hat{\beta}_{CGMM}) \end{aligned}$$

Now we have all the ingredients to define the distance statistic D, defined as

$$D \equiv J(\hat{\beta}_{CGMM}) - J(\hat{\beta}_{GMM})$$

Note that $D \geq 0$ by construction. This is because $J(\hat{\beta}_{CGMM})$ is from the minimization problem that is constrained and the other is an unconstrained problem, which usually results in lower values.

As we did for hypothesis testing on Wald test statistic, we can use D for hypothesis tests of the following setup

$$H_0 : r(\beta) = \theta, \;\; H_1 : r(\beta) \neq \theta$$

Under H_0, D converges in distribution to χ^2_q. We can find a critical value c s.t. $\alpha = 1 - F(c)$, where F is the CDF for χ^2_q and reject H_0 if $D > c$. The idea is that if H_0 is true, then imposing the restriction does not alter the moment equations greatly, making it a sensible restriction. Otherwise, the moment equation changes greatly, making it unreasonable constraint.

In fact, we can show that when $r(\beta)$ is linear, D becomes identical to the Wald Test Statistic W. We have shown what $\hat{\beta}_{CGMM}$ is. Then we can write

$$\begin{aligned} \bar{g}_n(\hat{\beta}_{CGMM}) &= \frac{Z'y - Z'X\hat{\beta}_{CGMM}}{n} \\ &= \frac{\overbrace{Z'y - Z'X(\hat{\beta}_{GMM}}^{=\bar{g}_n(\hat{\beta}_{GMM})} - \overbrace{(X'ZWZ'X)^{-1}R(R'(X'ZWZ'X)^{-1}R)^{-1}(R'\hat{\beta}_{GMM} - c))}^{=\hat{\beta}_{GMM} - \hat{\beta}_{CGMM}}}{n} \\ &= \bar{g}_n(\hat{\beta}_{GMM}) + \frac{Z'X(\hat{\beta}_{GMM} - \hat{\beta}_{CGMM})}{n} \end{aligned}$$

Thus,

$$\begin{aligned}
J(\hat{\beta}_{CGMM}) &= n\left(\bar{g}_n(\hat{\beta}_{GMM}) + \frac{Z'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})}{n}\right)' W \left(\bar{g}_n(\hat{\beta}_{GMM}) + \frac{Z'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})}{n}\right) \\
&= n\bar{g}_n(\hat{\beta}_{GMM})'W\bar{g}_n(\hat{\beta}_{GMM}) + \bar{g}_n(\hat{\beta}_{GMM})'WZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM}) \\
&+ (\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})'X'ZW\bar{g}_n(\hat{\beta}_{GMM}) + \frac{1}{n}(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})'X'ZWX'Z(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})
\end{aligned}$$

where $\bar{g}_n(\hat{\beta}_{GMM})'WZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})$ terms can be erased because

$$\begin{aligned}
\bar{g}_n(\hat{\beta}_{GMM})'WZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM}) &= \frac{y'ZWZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM}) - \hat{\beta}'_{GMM}X'ZWZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})}{n} \\
&= \frac{y'ZWZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM}) - y'ZWZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})}{n} = 0
\end{aligned}$$

since $\hat{\beta}_{GMM} = (X'ZWZ'X)^{-1}X'ZWZ'y$. Therefore,

$$J(\hat{\beta}_{CGMM}) = J(\hat{\beta}_{GMM}) + \frac{1}{n}(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})'X'ZWX'Z(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})$$

So $J(\hat{\beta}_{CGMM}) - J(\hat{\beta}_{GMM})$ becomes a Wald Test Statistic.

1.4 Overidentification Tests

In this section, we generalize the overidentification test we learned in 2SLS setup to a GMM setting. It is generalized in the sense that we can allow for heteroskedasticity whereas for 2SLS overidentification test, we relied on the assumption of homoskedasticity. If $\dim(g_i) = l > k = \dim(\beta)$, it is possible that there exists no β s.t. $E[g(w_i, \beta)] = 0$ is satisfied. Therefore, the overidentifying restrictions become testable. Effectively, we are testing the hypothesis

$$H_0 : E[g(w_i,\beta)] = 0 \text{ vs. } H_1 : E[g(w_i,\beta)] \neq 0$$

Let β_0 be the true value for the parameter of interest. Then $\bar{g}_n(\beta_0) = \frac{Z'y - Z'X\beta_0}{n} = \frac{Z'e}{n}$ has a limiting distribution characterized by

$$\sqrt{n}\bar{g}_n(\beta_0) \xrightarrow{d} N(0,\Omega), \quad \Omega \in \mathbb{R}^{l\times l}$$

and thus $J(\beta_0) \xrightarrow{d} \chi^2_l$. Since we are trying to estimate what β_0 is, we need to use a GMM estimator to build our test statistic $J = J(\hat{\beta}_{GMM})$, which has a limiting distribution χ^2_{l-k} under H_0. Like before, we can find c s.t. $\alpha = 1 - F(c)$, and then reject H_0 when $J > c$. However, it should be noted that when the H_0 is rejected, we only know that some moment condition is violated. We cannot pick out which. Nevertheless, rejection of the H_0 is a bad sign.

We can apply overidentification test on a (strict) subset of instruments whose validity is uncertain. To do this, we can partition z_i into two sets - $z_{ai} \in \mathbb{R}^{l_a}$ and $z_{bi} \in \mathbb{R}^{l_b}$. We are uncertain about z_{bi} and want to test

$$H_0 : E(z_{bi}e_i) = 0, \text{ vs. } H_1 : E(z_{bi}e_i) \neq 0$$

We can construct the test statistic in a following manner. First, estimate the model by the efficient GMM with only the z_{ai} set of instruments and obtain the GMM criterion. This will be denoted as J_a. Then, estimate the model with the full set of instruments and obtain a separate GMM criterion, denoted as $J_{a,b}$. Then, create a test statistic

$$C = J_{a,b} - J_a \xrightarrow{d} \chi^2_{l-l_a=l_b}$$

The idea is similar to that of a distance test. If the C value is low, then the moment condition for the larger set of instruments are not too different from the moment condition from the smaller (but surer) set of instruments. Then, we find a critical value c for a significance level α and reject the null hypothesis if $C > c$.

One example of a subset overidentification test is an endogeneity test. Assume a following data generating process

$$y_i = x'_{1i}\beta_1 + x'_{2i}\beta_2 + e_i$$

where x_{1i} is exogenous but x_{2i} may not be. Then, we want to test

$$H_0 : E(x_{2i}e_i) = 0, \text{ vs. } H_1 : E(x_{2i}e_i) \neq 0$$

Let $z_i = (x_{1i} \;\; z_{2i})' \in \mathbb{R}^l$ and assume that $E(z_ie_i) = 0$. Then we can create the test statistics as follows. First, estimate the efficient GMM by using (x_{1i}, z_{2i}) to instrument (x_{1i}, x_{2i}). Then obtain the GMM criterion $\tilde{J}$. Next, work with a larger set by using (x_{1i}, x_{2i}, z_{2i}) to instrument (x_{1i}, x_{2i}). After this, we can get the GMM criterion $\widehat{J}$. The relevant test statistic is

$$C = \widehat{J} - \tilde{J} \xrightarrow{d} \chi^2_{k_2}$$

As before, we find a relevant critical value c for a significance level α and reject H_0 of exogeneity for x_2 if $C > c$.

Again, the idea is as follows. We know that x_{1i} and z_{2i} are exogenous. If x_{2i} is also exogenous, then the GMM criterion obtained by using all three should not be too different from GMM criterion from variables x_{1i} and z_{2i}[1]. If the difference between the two GMM criterion is sufficiently large, we have a reason to suspect that, given that x_{1i} and z_{2i} is exogenous for sure, that the moment condition involving x_{2i} variables are vastly different from moment conditions from exogenous variables. Thus, x_{2i} could be endogenous.

1.5 Conditional Moment Conditions

In some cases, the moment condition for the data generating process can be expressed in terms of conditional moments. Assume that the data generating process and conditional moment condition of interest is

$$y_i = m(x_i, \beta^0) + e_i \quad E(e_i(\beta)|z_i) = 0$$

where β^0 is the true value of β, z_i is the instrument and $m(x_i, \beta)$ can be either linear or nonlinear. Usually, these conditions are more stronger in the sense these sets of conditions imply that the unconditional moment conditions assumed throughout the discussion holds. In fact, we can show that if $E(e_i|z_i) = 0$, then $E(h(z_i)e_i) = 0$ for measurable function $h(\cdot)$. The other way may not hold.

Example 1.1. *Suppose that $E(z_ie_i) = 0$, then is it the case that $E(e_i|z_i) = 0$ hold? There is this counterexample: Let z_i and e_i be characterized as*

$$z_i = \begin{cases} 1 & \Pr(z_i = 1) = 1/2 \\ 2 & \Pr(z_i = 2) = 1/2 \end{cases}, \quad e_i = \begin{cases} 1 & (\textit{if } z_i = 1) \\ -1/2 & (\textit{if } z_i = 2) \end{cases}$$

Then $E[z_ie_i]$ is $\frac{1}{2}(1 \times 1) + \frac{1}{2}(2 \times (-1/2)) = \frac{1}{2} - \frac{1}{2} = 0$. As for $E[e_i|z_i]$, we get $E[e_i|z_i = 1] = 1$ and $E[e_i|z_i = 2] = -1/2$. Therefore, the conditional mean is not 0.

So how do we solve when given conditional moments? One is to use the idea that $E[e_i|z_i] = 0 \implies E[h(z_i)e_i] = 0$ to construct some number of instruments and solve a

[1] It can be slightly different, since the weighting matrix used in calculating the GMM criterion is usually different. This can lead to some discrepancies even if H_0 turns out to be true.

GMM with unconditional moments[2]. The other is to construct an optimal instrument as Chamberlain (1987) found. For this, we need to define these equations

$$R(z_i) = E\left[\frac{\partial e_i(\beta^0)}{\partial \beta} \mid z_i\right] \in \mathbb{R}^k$$
$$\sigma^2(z_i) = E[e_i(\beta^0)^2|z_i]$$

Then, the optimal instrument is defined by

$$A_i = -\frac{R(z_i)}{\sigma^2(z_i)} \in \mathbb{R}^k$$

yielding the optimal moment

$$g_i^*(\beta) = A_i e_i(\beta)$$

Thus, by working with the $E[g_i^*(\beta)] = 0$ gives us the best GMM estimator possible (best being the lowest possible variance). In class, we have shown that if we actually knew the true σ^2, this gets us the GLS estimator.

2 Panel Data

2.1 Basic Framework

To motivate the discussion going forward, consider the following setting. Let y and $x = (x_1, x_2, ..., x_k)$ be the observable factors. Denote α as an unobservable random variable that is incorporated into the data generating process additively. Then, we can write $E[y|x, \alpha]$ as

$$E[y|x, \alpha] = x'\beta + \alpha$$

We are interested in estimating β. To see whether we can find a consistent estimator for β, we need to see how α is correlated with x. If α is independent from x, the it is not different from the idiosyncratic error and β is consistently estimable. However, in most cases, α could be correlated with x. If this is the case, then we cannot find a consistent estimator for β unless we find a perfect proxy for α. However, if α is fixed across time for an individual and we have access to panel data, we can address this issue. Most discussion of panel data in class will focus on α representing an unobservable trait inherent for individuals.

[2] There is no consensus on the optimal number of optimal instruments yet

In a panel data setting, the data now has two dimensions - dimensions across different unit of observation i and across time t. In maths,

$$(y_{it}, x_{it}) \text{ where } i = 1, ..., n, \text{ and } t = 1, ..., T$$

For our discussion, we will fix T and let the (y_{it}, x_{it}) be IID across $i = 1, ..., n$. Unless otherwise state, the asymptotics will be carried out by extending n to infinity.

To be more concrete, we can write the data generating process as

$$y_{it} = x'_{it}\beta + \alpha_i + u_{it} \text{ }^3$$

where x_{it} is an observable variable that varies across individuals and time periods. α_i is the **individual (fixed) effect** that is unobservable. As such, we can define $v_{it} = \alpha_i + u_{it}$ and rewrite the data generating process as

$$y_{it} = x'_{it}\beta + v_{it}$$

Depending on whether α_i is correlated with x_{it} or not, we can categorize α_i into the following

- **Random Effects**: There is no correlation between the observables and α_i
- **Fixed Effects**: The correlation between x_{it} and α_i is nonzero.

In case you are looking into some old textbooks, the categorization is slightly different. If α_i is considered to be a parameter to be estimated, the old textbooks refers to this as fixed effects. If α_i is a random variable, it was called a random effect. Note that in the above discussion, α_i is random variable in both fixed and random effects. You should see that what matters now is whether α_i is correlated with x_{it} or not.

2.2 Estimation

Before moving further, the following assumption is required to show whether the panel estimates are consistent or not

[3] Note that we can be interested in the time fixed effect, usually denoted as γ_t. The analysis of such effects is largely similar to individual fixed effects.

Assumption 2.1 (Strict Exogeneity). *We say that the regressor x_{it} is **strictly exogenous** with respect to u_{it} if*

$$E[y_{it}|x_{i1},..,x_{iT},\alpha_i] = E[y_{it}|x_{it},\alpha_i]$$

which boils down to

$$E[u_{it}|x_{i1},..,x_{iT},\alpha_i] = E[u_{it}|x_{it},\alpha_i] = 0$$

In words, strict exogeneity implies that once x_{it},α_i is controlled for, $x_{is}, s \neq t$ has no partial effect on y_{it}. The above condition also implies that

$$E[x_{is}u_{it}] = 0 \; (s,t) = 1,..,T$$

while this itself is a weaker condition that $E[u_{it}|x_{i1},..,x_{iT},\alpha_i] = 0$, it is sufficient for most of the consistency analysis. However, it is stronger than the exogeneity assumptions of the previous chapters. Previously it was enough for the regressors to be exogenous with the contemporaneous e_{it}. Now we assume x_{it} to be exogenous with all error terms. [4]

Also note that this condition is silent about the correlation between x_{it} and α_i.

2.2.1 Pooled OLS (POLS)

Pooled OLS estimation practically ignores the panel structure of the data and involves applying a typical OLS estimation on the data generating process with time and individual dimension. Given that our model is

$$y_{it} = x_{it}'\beta + \underbrace{\alpha_i + u_{it}}_{=v_{it}}$$

The estimator can be written as

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

Under the situation where $cov(x_{it}.\alpha_i) \neq 0$, we can see that $\hat{\beta}_{POLS}$ is inconsistent. Note that we can re-write the POLS estimator in terms of the true β

$$\hat{\beta}_{POLS} - \beta = \left(\sum_{i=1}^{n}\sum_{t=1}^{T} x_{it}x_{it}'\right)^{-1} \sum_{i=1}^{n}\sum_{t=1}^{T} x_{it}v_{it}$$

[4] Yes, you may think that this assumption is extreme. Later in the panel data discussion, we will talk about weaker version of this assumption - sequential exogeneity.

The trick is to show that $E[x_{it}v_{it}] \neq 0$ for $t = 1, .., T$. Since I can write $E[x_{it}(\alpha_i + u_{it})]$ instead, it can be seen that unless $cov(x_{it}, \alpha_i) = 0$, the $E[x_{it}\alpha_i]$ term will remain. Thus, POLS estimator is not consistent.

2.2.2 First Difference (FD)

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

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

where $\Delta y_{it} = y_{i,t} - y_{i,t-1}$ and similarly for other variables. Notice that since α_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 can show that this is a consistent estimator. Write

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

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

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

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. This would effectively rule out time-constant regressors.

2.2.3 Within Estimators (WE)

This shares a similar idea with First Difference in the sense that it attempts to weed out α_i. This approach uses a different method. I will define new notations that involve variables that are averaged across time. Write

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$$

and similarly for other variables. By average over time across all variables, we can get a cross-sectional equation

$$\bar{y}_i = \bar{x}_i'\beta + \alpha_i + \bar{u}_i$$

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

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

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

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

To show consistency, rewrite the above as

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

$E[\tilde{x}_{it}\tilde{u}_{it}]$ can be written as

$$E[\tilde{x}_{it}\tilde{u}_{it}] = E[x_{it}u_{it}] - E[x_{it}\bar{u}_i] - E[\bar{x}_i u_{it}] + E[\bar{x}_i\bar{u}_i]$$

Since $\bar{x}_i, \bar{u}_i$ 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 $E\left[\tilde{x}_{it}\tilde{x}_{it}'\right]$ be a full column rank for its inverse to be defined. Time-constant regressors are ruled out.

2.2.4 Least Square Dummy Variables (LSDV)

This comes from viewing α_i as a parameter to be estimated. 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[5]. Therefore, we work with

$$y_{it} = x_{it}'\beta + D1_i\alpha_1 + ... + D(n-1)_i\alpha_{n-1} + u_{it}$$

What is going on here is that each individual has his/her own constant term. 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 $\beta_0 + \alpha_k$.

2.2.5 Some Interesting Topics

- **WE vs FD**: When $T = 2$, it can be shown that they are numerically equal. I start with the WE, written as

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

  Each term can be replaced by

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

  and

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

  By combining the two, I get $\left(\sum_{i=1}^{n}\Delta x_{i2}\Delta x_{i2}'\right)^{-1}\sum_{i=1}^{n}\Delta x_{i2}\Delta y_{i2}$, which is equivalent to FD when $T = 2$.

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

When $T \geq 3$, they are no longer equal. So it begs the question as to which estimator to use. This depends on the structure of the error terms. If u_{it} is free from serial correlation (or IID), then taking a first difference would introduce serial correlation. This is because

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

As such, first difference in this situation suffers from inconsistency problems arising due to serial correlation.

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

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

If we use a first difference estimator here, we get to obtain the most efficient estimator.

- **WE=LSDV** To show that they are numerically equal, it helps to know how these estimators can be written in terms of matrices and understand what Kronecker product is. Note the following

Definition 2.1 (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)
- $(A \otimes B)' = (A' \otimes B')$

I will first write WE in terms of matrices. for each individual i, we can stack up total of T observations and write

$$\mathbf{y}_i = \mathbf{X}_i\beta + \alpha_i 1_T + \mathbf{u}_i \in \mathbb{R}^T$$

where 1_T is the $T \times 1$ vector of 1′s as elements. Now I define $Q_T \equiv I_T - 1_T(1_T'1_T)^{-1}1_T'$ Note that $(1_T'1_T)^{-1} = T^{-1}$ and $1_T1_T'$ is the T-dimensional square matrix of 1′s as elements. This is symmetric and idempotent. Also note that

$$Q_T1_T = 0, Q_T\mathbf{y}_i = \mathbf{y}_i - \frac{1}{T}\begin{pmatrix}\sum_t y_{it}\\ ...\\ \sum_t y_{it}\end{pmatrix} = \begin{pmatrix}y_{i1} - \frac{1}{T}\sum_t y_{it}\\ ...\\ y_{iT} - \frac{1}{T}\sum_t y_{it}\end{pmatrix} = \tilde{\mathbf{y}}_i, Q_T\mathbf{X}_i = \tilde{\mathbf{X}}_i, Q_T\mathbf{u}_i = \tilde{\mathbf{u}}_i$$

This shows that pre-multiplying Q_T to $\mathbf{y}_i = \mathbf{X}_i\beta + \alpha_i 1_T + \mathbf{u}_i \in \mathbb{R}^T$ gives us the demeaned version of the original data generating process. So we need to take OLS on

$$Q_T\mathbf{y}_i = Q_T\mathbf{X}_i\beta + Q_T\mathbf{u}_i$$
$$\implies \hat{\beta}_{WE} = \left(\sum_{i=1}^n \mathbf{X}_i'Q_T\mathbf{X}_i\right)^{-1}\sum_{i=1}^n \mathbf{X}_i'Q_T\mathbf{y}_i$$

We can characterize the LSDV estimator from the same original data generating process. What we need to do is to stack $\mathbf{y}_i$ and $\mathbf{u}_i$ for each i and obtain a nT dimensional vector $\mathbf{y}$ and $\mathbf{u}$. Similarly, stack each $\mathbf{X}_i$ to get an $nT \times k$ matrix $\mathbf{X}$. Since each α_i exists for each individual i, the idea here is to use a Kronecker product and express the individual fixed effect as

$$\begin{pmatrix}1_T & ... & 0\\ ... & ... & ...\\ 0 & ... & 1_T\end{pmatrix}\begin{pmatrix}\alpha_1\\ ...\\ \alpha_n\end{pmatrix} = (I_n \otimes 1_T)\alpha \in \mathbb{R}^{nT\times n} \times \mathbb{R}^{n\times 1} = \mathbb{R}^{nT\times 1}$$

Combine what we know to get

$$\mathbf{y} = \mathbf{X}\beta + (\underbrace{I_n \otimes 1_T}_{=D})\alpha + \mathbf{u}$$

We then use the Frisch-Waugh-Lovell theorem (Chapter 3, Hansen (2019)) to get

$$\hat{\beta}_{LSDV} = (\mathbf{X}'M_D\mathbf{X})^{-1}(\mathbf{X}'M_D\mathbf{y})$$

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

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

Thus, $M_D = I_{nT} - I_n \otimes 1_T(1_T'1_T)^{-1}1_T' = I_n \otimes I_T - I_n \otimes 1_T(1_T'1_T)^{-1}1_T'$. By using the first property mentioned with the Kronecker product, this is equal to $I_n \otimes (I_T - 1_T(1_T'1_T)^{-1}1_T') = I_n \otimes Q_T$. So $\hat{\beta}_{LSDV}$ is equal to

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

Since $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ ... \\ \mathbf{X}_n \end{pmatrix}$, $I_n \otimes Q_T = \begin{pmatrix} Q_T & 0 & 0 \\ ... & ... & ... \\ 0 & 0 & Q_T \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ ... \\ \mathbf{y}_n \end{pmatrix}$, we get

$$\hat{\beta}_{LSDV} = \left(\sum_{i=1}^{n} \mathbf{X}_i' Q_T \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i' Q_T \mathbf{y}_i = \hat{\beta}_{WE}$$

and we are done.