Econometrics II TA Session *

Wang Xin

Nov 29, 2023

From next week, the TA session will be given at 15:10 every Monday.

1 Supplement of Panel Data

1.1 Panel Data

A panel data set, while having both a cross-sectional and a time series dimension, differs in some important respects from an independently pooled cross section. To collect panel data, also called longitudinal data, we follow (or attempt to follow) the same individuals about the same variables across time.

In our class, we only consider balanced panels, which means that each individual in the data set is observed the same number of times, usually denoted T.

Suppose that for each cross section unit, we observe data on the same set of variables for n periods. Let $\mathbf{X}_{it}$ a $1 \times K$ vector, and $\boldsymbol{\beta}$ a $K \times 1$ vector. The model in the population is

$$y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + u_{it}$$

where y_{it} and u_{it} are scalars, $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T$. This model is called linear panel data model.

Before we start to perform estimation, we need to know some assumptions at first:

- Contemporaneous exogeneity assumption: $E[u_{it}|\mathbf{X}_{it}] = 0, \ \forall\ i, t$
- Strict exogeneity assumption: $E[u_{it}|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT}] = 0$

Contemporaneous exogeneity assumption places no restrictions on the relationship between $\mathbf{X}_{is}$ and u_{it} for $s \neq t$. Strict exogeneity assumption implies that each u_{it} is uncorrelated with the explanatory variables in all time periods.

Reformulate the panel model as the following regression system:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i \tag{1}$$

where $\mathbf{X}_i = (\mathbf{X}'_{i1}, \mathbf{X}'_{i2}, \ldots, \mathbf{X}'_{iT})'$ is a $T \times K$ matrix, $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$ and $\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iT})'$ are $T \times 1$ vectors.

*Cited from documents by Professor Poignard.

1.2 Pooled OLS

Still consider the model in Eq(1), before we can apply OLS estimation, there are two assumptions we need to establish:

- $\forall\ i,t\ E[\mathbf{X}'_{it}u_{it}] = 0$. Under this assumption, we have the orthogonality condition $\forall\ i,t\ E[\mathbf{X}'_{it}(y_{it} - \mathbf{X}_{it}\boldsymbol{\beta})] = 0$.
- $E[\mathbf{X}'_i\mathbf{X}_i] \succ 0$, which means a positive definite matrix.

Then the OLS estimator is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n}\mathbf{X}'_i\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}'_i\mathbf{y}_i\right) = \left(\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{X}'_{it}\mathbf{X}_{it}\right)^{-1}\left(\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{X}'_{it}\mathbf{y}_{it}\right)$$

This estimator in the context of panel data is called pooled OLS.

1.3 Generalized Least Squares Estimation

In above model, we actually didn't take heteroskedasticity with respect to errors into account, which is quite common in panel data. To obtain consistent estimaors, we introduce the generalized least squares (GLS) analysis.

Define the variance of error term as a $T \times T$ matrix $\Omega = E[\mathbf{u}_i\mathbf{u}'_i]$. Then the assumptions can be rewritten as:

- $E[\mathbf{X}_i \otimes \mathbf{u}_i] = 0$. This assumption is actually more strict than the assumption for Pooled OLS. This assumption implies that $E[\mathbf{X}'_i\Omega^{-1}\mathbf{u}_i] = 0$.
- $\Omega \succ 0$ and $E[\mathbf{X}'_i\Omega^{-1}\mathbf{X}_i] \succ 0$

By GLS method we obtain,

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\sum_{i=1}^{n}\mathbf{X}'_i\Omega^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}'_i\Omega^{-1}\mathbf{y}_i\right)$$

In full matrix notation, it is given by

$$\hat{\boldsymbol{\beta}}_{GLS} = \left\{\mathbf{X}'(\mathbf{I}_n \otimes \Omega^{-1})\mathbf{X}\right\}^{-1}\left\{\mathbf{X}'(\mathbf{I}_n \otimes \Omega^{-1})\mathbf{y}\right\}$$

2 Review of Individual effect

We previously assumed the so-called exogeneity assumption. This hypothesis is actually too strong for certain panel data. In fact, a primary motivation for using panel data is to solve the **omitted variables problem**.

Consider y_{it} and $\mathbf{X}_{it}$ be observable random variables and v_i an **unobservable** random variable.

$$y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + v_i + u_{it} \tag{2}$$

There are K regressors in $\mathbf{X}_{it}$, not including a constant term. The heterogeneity, or **individual effect** is a scalar, v_i, where v_i contains a constant term and a set of individual or group-specific variables, which may be observed, such as race, sex, location, and so on, or unobserved, such as family specific characteristics, individual heterogeneity in skill or preferences, and so on, all of which are taken to be constant over time t.

The key assumption for individual effect, both fixed effect and random effect, is the strict exogeneity assumption on the explanatory variables:

- **A.1** $E[u_{it}|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT}, v_i] = 0$

2.1 Fixed effect model

In the fixed effect framework, v_i is not treated as non-random; rather, it means that one is allowing for arbitrary dependence between the unobserved effect v_i and the observed explanatory random variables. In short, v_i is correlated with $\mathbf{X}_{it}$, then the least squares estimator of $\boldsymbol{\beta}$ is biased and inconsistent as a consequence of an omitted variable.

An alternative notation of OLS estimation of fixed effect model is displayed below, if you find the Kronecker product difficult.

Rewrite the model (2) as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{1}_T v_i + \mathbf{u}_i \tag{3}$$

with where $\mathbf{X}_i = (\mathbf{X}'_{i1}, \mathbf{X}'_{i2}, \ldots, \mathbf{X}'_{iT})'$ is a $T \times K$ matrix, $\mathbf{X}_{it}$ is a $1 \times K$ vector, $\mathbf{y}_i = (\mathbf{y}'_{i1}, \mathbf{y}'_{i2}, \ldots, \mathbf{y}'_{iT})'$, $\mathbf{u}_i = (\mathbf{u}'_{i1}, \mathbf{u}'_{i2}, \ldots, \mathbf{u}'_{iT})'$ and $\mathbf{1}_T = (1, 1, \ldots, 1)'$ are $T \times 1$ vectors.

The main idea for estimating $\boldsymbol{\beta}$ under **A.1** is to transform the equations to eliminate the unobserved effect v_i. Average the model (3) over time, we have

$$\bar{\mathbf{y}}_i = \bar{\mathbf{X}}_i\boldsymbol{\beta} + \mathbf{1}_T v_i + \bar{\mathbf{u}}_i \tag{4}$$

where $\bar{\mathbf{y}}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\bar{\mathbf{X}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_{it}$ and $\bar{\mathbf{u}}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$.

v_i can be eliminated by subtracting Eq(4) from Eq(3), and remains

$$\ddot{\mathbf{y}}_i = \ddot{\mathbf{X}}_i\boldsymbol{\beta} + \ddot{\mathbf{u}}_i$$

where $\ddot{\mathbf{y}}_i = \mathbf{y}_i - \bar{\mathbf{y}}_i$, $\ddot{\mathbf{X}}_i = \mathbf{X}_i - \bar{\mathbf{X}}_i$ and $\ddot{\mathbf{u}}_i = \mathbf{u}_i - \bar{\mathbf{u}}_i$.

In order to ensure that asymptotically the FE estimator is well behaved, we make a standard rank condition:

- **FE2** $rank\left(\sum_{t=1}^{T} E[\ddot{\mathbf{X}}_{it}'\ddot{\mathbf{X}}_{it}]\right) = rank\left(E[\ddot{\mathbf{X}}_{i}'\ddot{\mathbf{X}}_{i}]\right) = K$

The FE estimator is the pooled OLS estimator from the regression $\ddot{\mathbf{y}}_i$ on $\ddot{\mathbf{X}}_i$, which is

$$\hat{\boldsymbol{\beta}}_{FE} = \left(\sum_{i=1}^{n} \ddot{\mathbf{X}}_{i}'\ddot{\mathbf{X}}_{i}\right)^{-1}\left(\sum_{i=1}^{n} \ddot{\mathbf{X}}_{i}'\ddot{\mathbf{y}}_{i}\right) = \left(\sum_{i=1}^{n}\sum_{t=1}^{T} \ddot{\mathbf{X}}_{it}'\ddot{\mathbf{X}}_{it}\right)^{-1}\left(\sum_{i=1}^{n}\sum_{t=1}^{T} \ddot{\mathbf{X}}_{it}'\ddot{\mathbf{y}}_{it}\right)$$

2.2 Random effect Model

A random effect is synonymous with zero correlation between the observed explanatory random variables and the unobserved effect

$$Cov(\mathbf{X}_{it}, v_i) = 0,$$

Along with **A.1**, the assumption can be restate as

- **RE1** $E[u_{it}|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT}, v_i] = 0$ and $E[v_i|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT}] = 0$

In the random effect method, the variable v_i can be put into the error term, which is

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \text{ with } \boldsymbol{\epsilon}_i = \mathbf{1}_T v_i + \mathbf{u}_i$$

and $E[\epsilon_{it}|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT}] = 0$

We define the variance covariance matrix as $\Omega = E[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i'] \succ 0$

In addition, we assume

- **RE2** $rank\left(E[\mathbf{X}_i'\Omega^{-1}\mathbf{X}_i]\right) = K$
- **RE3** $E[\mathbf{u}_i\mathbf{u}_i'|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT}, v_i] = \sigma_u^2\mathbf{I}_T$ and $E[v_i^2|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT}] = \sigma_v^2$

Remember that v_i is a scalar and constant over time t, then we can derive the $T \times T$ variance covariance matrix Ω as

$$\Omega = \begin{pmatrix} \sigma_v^2 + \sigma_u^2 & \sigma_v^2 & \ldots & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 + \sigma_u^2 & \ldots & \vdots \\ \vdots & \vdots & \ddots & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 & \ldots & \sigma_v^2 + \sigma_u^2 \end{pmatrix}$$

Except for the Maximum Likelihood Estimation, which is shown on Professor's class, **Feasible GLS estimation** can also be applied to obtain the estimator.

Remember in Section 1.3, we derive the GLS estimator as

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\sum_{i=1}^{n} \mathbf{X}_i'\Omega^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{X}_i'\Omega^{-1}\mathbf{y}_i\right)$$

However, since v_i is an unobservable variable, we will not obtain the specific value of Ω. Therefore, we need to estimate Ω first.

To do so, we estimate the regression $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$ by OLS and obtain $\hat{\boldsymbol{\beta}}_{OLS}$ together with

$$\hat{\epsilon}_{it} = y_{it} - \mathbf{X}_{it}\hat{\boldsymbol{\beta}}_{OLS}$$

Then a consistent estimator of $\hat{\sigma_\epsilon}$ is

$$\hat{\sigma_\epsilon}^2 = \frac{1}{nT-K}\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{\epsilon}_{it}^2$$

Pay attention that $\hat{\sigma_\epsilon}^2$ is the estimator of the diagonal element in Ω, which is $\sigma_v^2 + \sigma_u^2$. Then we still need to estimate $\hat{\sigma_v}^2$.

As for a consistent estimator of σ_v^2, let us start from its definition:

$$\sigma_v^2 = E[\epsilon_{it}\epsilon_{is}],\ t \neq s$$

This means for each t, there are $T(T-1)/2$ redundant error products that can be used to estimate σ_v^2.

$$\begin{aligned} E[\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\epsilon_{it}\epsilon_{is}] &= \sum_{t=1}^{T-1}\sum_{s=t+1}^{T}E[\epsilon_{it}\epsilon_{is}] = \sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\sigma_v^2 \\ &= \sum_{t=1}^{T-1}\sigma_v^2(T-t) \\ &= \sigma_v^2 T(T-1)/2 \end{aligned}$$

Thus a consistent estimator is

$$\hat{\sigma_v}^2 = \frac{1}{nT(T-1)/2-K}\sum_{i=1}^{n}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\hat{\epsilon}_{it}\hat{\epsilon}_{is}$$

And you can even derive the $\hat{\sigma_u}^2$ by $\hat{\sigma_v}^2 - \hat{\sigma_v}^2$.

$\hat{\Omega}$ is estimated, thus the FGLS estimator can be deducted as

$$\hat{\boldsymbol{\beta}}_{FGLS} = \left(\sum_{i=1}^{n}\mathbf{X}_i'\hat{\Omega}^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i'\hat{\Omega}^{-1}\mathbf{y}_i\right)$$