TA Session 5

Bruno Amat

November 6, 2024

Outline

1. Panel Data

Panel Data

INTRODUCTION

A Panel Data can be defined as a dataset which we can observe the same individual/cities/firms for serveral periods. Hence, we often use the following notation

- i represents the individual of the observation
- t represents the period of the observation

Let Y represent the income. Hence Y_{it} represents the income of agent i during time t

INTRODUCTION

We have different types of Panel Data. In the TA session I will focus on the balanced and unbalanced panels

- **Balanced:** We have observations for all individuals during all periods
- **Unbalanced:** We are not able to identify the observations of one agent for all periods

The unbalanced panel might generate the **attrition** problem, which the individual does not appear anymore.

MODEL

We want to estimate β such that

$$Y_{it} = X_{it}\beta + u_{it}$$

We can make three different hypothesis about the error term:

- $E[u_{it}|X_{it}]$ = 0 (pooled estimation)
- $u_{it} = \alpha_i + \epsilon_{it}$ (one-way error component)
- $u_{it} = \alpha_i + \lambda_t + \epsilon_{it}$ (two-way error component)

Random and Fixed Effects

Using the two-way error component, we can see the parameters α_i and λ_t in two different ways:

- **Fixed Effects:** Parameters
- **Random Effects:** Random Variables

Fixed Effects

We want to estimate the following model

$$Y_{it} = X_{it}\beta + \alpha_i + \epsilon_{it}$$

We can estimate β using the following estimation

$$\tilde{Y_{it}} = \tilde{X_{it}}\beta + \tilde{\epsilon_{it}}$$

Where $\tilde{X_{it}} = \tilde{X_{it}} - \bar{Y_i}$ and $\bar{Y_i} = T^{-1}\sum Y_{it}$. By using this approach we can cancel the α_i part from the error term

Fixed Effects

If $T = 2$ we can also use the first difference estimator

$$\Delta Y_{it} = \Delta X_{it}\beta + \Delta\epsilon_{it}$$

where $\Delta Y_{it} = Y_{it} - Y_{it-1}$. If $T > 2$ the estimator is different from the fixed effect.

Fixed Effects

We can also estimate the Fixed Effect estimator by using a LSDV approach

$$Y_{it} = X_{it}\beta + D_i'\alpha + \epsilon_{it}$$

where $D_i = 1$ for individual i, else $D_i = 0$.
Using the FWL theorem we obtain $\beta_{LSDV} = \beta_{FE}$.

Random Effects

In this estimation we treat α_i as a random variable. Hence, we need some assumptions to perform the random effect estimation:

$$E[\alpha_i|X_i] = E[\epsilon_{it}|X_i] = 0$$
$$E[\alpha_i^2|X_i] = \sigma_\alpha^2,\ E[\alpha_i \epsilon_{it}|X_i] = 0$$
$$E[\epsilon_{it}\epsilon_{jt}|X_i] = 0,\ E[\epsilon_{it}^2|X_i] = \sigma_\epsilon^2.$$

Hence, we can use a GLS estimation to obtain the random effects parameter

$$\hat{\beta}_{RE} = \left(\sum_{i=1}^{N} X_i' \Omega_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' \Omega_i^{-1} Y_i\right)$$

HAUSMAN TEST

We can use the Hausman test to test the endogeneity of a variable.

Note that in the Fixed-Effect estimation we treat D_i as a relevant part of the error term that if it is not added to the model, we would have problems of omitted variables.

In the other hand, the Random Effects treat α_i as a random variable satisfying the exogeneity property $E[\alpha_i|X_i] = 0$.

Hence the Hausman test tests if D_i is exogenous, against the alternative of D_i being endogenous. The test statistic is given by:

$$\frac{(\hat{\beta_{FE}} - \hat{\beta_{RE}})^2}{\sigma_{\hat{\beta}_{FE}}^{\ 2} - \sigma_{\hat{\beta}_{FE}}^{\ 2}} \sim \chi^2_{k-1}$$

Time Effects

Consider the LSDV approach for the Fixed Effect.

Note that we are considering only the individual fixed effect.

However, time might also influence the dependent variable. Hence, we can use the following expression to control for time.

$$Y_{it} = X_{it}\beta + D_i'\alpha + +\gamma t + \epsilon_{it}$$

TIME EFFECTS

One limitation of the previous approach is that we are considering a linear time trend. Consequently, we are aggregating all periods into one variable, this might result in a strange parameter.

Consequently, a random shock like the pandemics influence the parameter in some point that its interpretation might not be precise. Hence, we can use **Time Fixed Effects** to account for the time effect and capture more heterogeneity.

In this formulation we add dummies for each period in our data base. Consequently, the model is given by:

$$Y_{it} = X_{it}\beta + D_i'\alpha + \gamma' D_t + \epsilon_{it}$$

Two Way Fixed Effect

Suppose we want estimate the following

$$Y_{it} = X_{it}\beta + u_{it}$$

Where $u_{it} = \lambda_t + \alpha_i + \epsilon_{it}$. Hence the TWFE estimator is given by:

$$\tilde{Y}_{it} = \tilde{X}_{it}\beta + \tilde{\epsilon}_{it}$$

where $\tilde{Y}_{it} = Y_{it} - \bar{Y}_i - \bar{Y}_t + \bar{Y}$

Dynamic Panel

We can also add lagged variables to our estimation using a panel dataset. For example

$$Y_{it} = \theta Y_{it-1} + X_{it}\beta + \alpha_i + \epsilon_{it}$$

However, using this approach we might generate problems of endogeneity. Hence, we need to come up with a Instrumental Variable approach and use the GMM to estimate the model.

Anderson and Hsiao (1982)

By using the first difference approach we obtain

$$\Delta Y_{it} = \theta \Delta Y_{it-1} + \Delta X_{it}\beta + \Delta\epsilon_{it}$$

Hence, they use ΔY_{it-2} or Y_{it-2} as an instrument for ΔY_{it-1}, since

$$Cov(\Delta Y_{it-2}, \Delta\epsilon_{it}) = Cov(Y_{it-2} - Y_{it-3}, \epsilon_{it} - \epsilon_{it-1}) = 0$$
$$Cov(\Delta Y_{it-2}, \Delta Y_{it-1}) = Cov(Y_{it-2} - Y_{it-3}, Y_{it-1} - Y_{it-2}) \neq 0$$

Anderson and Hsiao (1982)

Therefore, we can estimate the model by using the IV estimator:

$$(\hat{\gamma}, \hat{\beta}) = (Z'\tilde{X})^{-1}(Z'\tilde{Y})$$

where $\tilde{Y} = Y_{it} - Y_{it-1}$, $\tilde{X} = (\Delta Y_{it-1}, \Delta X'_{it})$, $Z = \Delta Y_{it-2}$

Arellano-Bond (1991)

The Arellano-Bond approach uses a GMM estimation with one moment equation for each period:

if T = 3:

$$\Delta Y_{i3} = \theta \Delta Y_{i2} + \Delta \epsilon_{i3}$$

Hence we have one moment equation and instrument ΔY_{i2} using Y_{i1}

Arellano-Bond (1991)

if T = 4:

$$\Delta Y_{i4} = \theta \Delta Y_{i3} + \Delta X_{i4}\beta + \Delta \epsilon_{i4}$$
$$\Delta Y_{i3} = \theta \Delta Y_{i2} + \Delta X_{i3}\beta + \Delta \epsilon_{i3}$$

Hence we have two moment equations, where we instrument ΔY_{i3} using Y_{i1}, Y_{i2} and instrument ΔY_{i2} using Y_{i1}

ARELLANO-BOND (1991)

Consequently, the moment conditions are given by:

$$E[Z_i(\epsilon_{it} - \epsilon_{it-1})] = 0$$

And

$$Z_i = \begin{bmatrix} Y_{i1} & 0 & ... & 0 \\ 0 & Y_{i1}Y_{i2} & ... & 0 \\ . & . & & \\ . & & . & \\ . & & & . \\ 0 & 0 & ... & Y_{i1}Y_{it-2} \end{bmatrix}$$

ARELLANO-BOND (1991)

We can estimate the model by minimizing:

$$\Delta\epsilon' Z A_N Z' \Delta\epsilon$$

Where A_N us the weighting matrix, that can be defined as:
two-step GMM:

$$A_N = \left(N^{-1} \sum_{i=1}^{N} Z_i \Delta\hat{\epsilon}_i \Delta\hat{\epsilon}_i' Z_i \right)^{-1}$$

ARELLANO-BOND (1991)

one-step GMM: Let H be a matrix with main diagonal equal to 2, -1 on the second main diagonal and 0 everywhere else.

$$A_N = \left(N^{-1} \sum_{i=1}^{N} Z_i H Z_i \right)^{-1}$$