Chapter 9

Panel Data

Economists mostly work with observational data. The data generation process is out of the researchers' control. If we only have a cross sectional dataset at hand, it is difficult to control heterogeneity among the individuals. On the other hand, panel data offers a chance to control heterogeneity of some particular forms.

A panel dataset tracks the same individuals across time $t = 1, \ldots, T$. We assume the observations are independent across $i = 1, \ldots, n$, while we allow some form of dependence within a group across $t = 1, \ldots, T$ for the same i. We maintain the linear equation

$$y_{it} = \beta_1 + x_{it}\beta_2 + u_{it},\ i = 1, \ldots, n; t = 1, \ldots, T \tag{9.1}$$

where $u_{it} = \alpha_i + \epsilon_{it}$ is called the *composite error*. Note that α_i is the time-invariant unobserved heterogeneity, while ϵ_{it} varies across individuals and time periods.

The most important techniques of panel data estimation are the fixed effect regression and the random effect regression. The asymptotic distributions of both estimators can be derived from knowledge about the OLS regression. In this sense, panel data estimation becomes applied examples of the theory that we have covered in this course. It highlights the fundamental role of linear regression theory in econometrics.

9.1 Fixed Effect

The unobservable individual-specific heterogeneity α_i is absorbed into the composite error $u_{it} = \alpha_i + \epsilon_{it}$. If $\text{cov}(\alpha_i, x_{it}) = 0$, the OLS is consistent; otherwise the consistency breaks down. The fixed effect model allows α_i and x_{it} to be arbitrarily correlated. The trick to regain consistency is to eliminate α_i.

This section develops the consistency and asymptotic distribution of the *within estimator*, the default fixed-effect (FE) estimator. The within estimator transforms the data by subtracting all the observable variables by the corresponding group means. Averaging the T equations of the original regression for the same i, we have

$$\overline{y}_i = \beta_1 + \overline{x}_i\beta_2 + \bar{u}_{it} = \beta_1 + \overline{x}_i\beta_2 + \alpha_i + \bar{\epsilon}_{it}. \tag{9.2}$$

where $\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$. Subtracting the averaged equation from the original equation gives

$$\tilde{y}_{it} = \tilde{x}_{it}\beta_2 + \tilde{\epsilon}_{it} \tag{9.3}$$

where $\tilde{y}_{it} = y_{it} - \overline{y}_i$. We then run OLS with the demeaned data, and obtain the within estimator

$$\widehat{\beta}_2^{FE} = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\tilde{y},$$

where $\tilde{y} = (y_{it})_{i,t}$ stacks all the nT observations into a vector, and similarly defined is $\tilde{X}$ as an $nT \times K$ matrix, where K is the dimension of β_2.

We know that OLS would be consistent if $E[\tilde{\epsilon}_{it}|\tilde{x}_{it}] = 0$. Below we provide a sufficient condition, which is often called *strict exogeneity*.

Assumption FE.1 $E[\epsilon_{it}|\alpha_i, \mathbf{x}_i] = 0$ where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iT})$.

Its strictness is relative to the contemporary exogeneity $E[\epsilon_{it}|\alpha_i, x_{it}] = 0$. FE.1 is more restrictive as it assumes that the error ϵ_{it} is mean independent of the past, present and future explanatory variables.

When we talk about the consistency in panel data, typically we are considering $n \to \infty$ while T stays fixed. This asymptotic framework is appropriate for panel datasets with many individuals but only a few time periods.

Proposition If FE.1 is satisfied, then $\widehat{\beta}_2^{FE}$ is consistent.

The variance estimation for the FE estimator is a little bit tricky. We assume a homoskedasitcity condition to simplify the calculation. Violation of this assumption changes the form of the asymptotic variance, but does not jeopardize the asymptotic normality.

Assumption FE.2 $\text{var}\left(\epsilon_i | \alpha_i, \mathbf{x}_i\right) = \sigma_\epsilon^2 I_T$ for all i.

Under FE.1 and FE.2, $\widehat{\sigma}_\epsilon^2 = \frac{1}{n(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T} \widehat{\tilde{\epsilon}}_{it}^2$ is a consistent estimator of σ_ϵ^2, where $\widehat{\tilde{\epsilon}} = \tilde{y}_{it} - \tilde{x}_{it} \widehat{\beta}_2^{FE}$. Note that the denominator is $n\left(T-1\right)$, not nT. The necessity of adjusting the degree of freedom can be easily seen from the FWL theorem: the FE estimator for the slope coefficient is numerical the same as its counterpart in the full regression with a dummy variable for each cross sectional unit.

If FE.1 and FE.2 are satisfied, then

$$\left(\widehat{\sigma}_\epsilon^2 \left(\tilde{X}'\tilde{X}\right)^{-1}\right)^{-1/2} \left(\widehat{\beta}_2^{FE} - \beta_2^0\right) \xrightarrow{d} N\left(0, I_K\right).$$

Proof. Let $M_\iota = I_T - \frac{1}{T}\iota_T \iota_T'$ be the within-group demeaner, and $M = I_n \otimes M_\iota$ ("$\otimes$" denotes the Kronecker product). The FE estimator can be explicitly written as

$$\widehat{\beta}_2^{FE} = \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'\tilde{Y} = \left(X'MX\right)^{-1} X'MY.$$

So

$$\sqrt{nT}\left(\widehat{\beta}_2^{FE} - \beta_2^0\right) = \left(\frac{X'MX}{nT}\right)^{-1} \frac{X'M\epsilon}{\sqrt{nT}} = \left(\frac{\tilde{X}'\tilde{X}}{nT}\right)^{-1} \frac{\tilde{X}'\epsilon}{\sqrt{nT}}$$

Since

$$\text{var}\left(\frac{\tilde{X}'\epsilon}{\sqrt{nT}} | X\right) = \frac{1}{nT} E\left(X'M\epsilon\epsilon'MX | X\right) = \frac{1}{nT} X'ME\left(\epsilon\epsilon' | X\right) MX = \left(\frac{\tilde{X}'\tilde{X}}{nT}\right)\sigma^2,$$

We apply a law of large numbers and conclude

$$\left(\tilde{X}'\tilde{X}\right)^{1/2}\left(\widehat{\beta}_2^{FE}-\beta_2^0\right)\xrightarrow{d}N\left(0,\sigma_\epsilon^2 I_K\right).$$

For simplicity, suppose we can direct observe $\tilde{\epsilon}_{it}$. Then

$$\begin{aligned}\frac{1}{n\left(T-1\right)}E\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{\epsilon}_{it}^2\right]&=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T-1}E\left[\epsilon_i' M_\iota \epsilon_i\right]\\&=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T-1}\mathrm{tr}\left(E\left[M_\iota E\left[\epsilon_i\epsilon_i'|\mathbf{x}_i\right]\right]\right)\\&=\frac{\sigma_\epsilon^2}{n}\sum_{i=1}^{n}\frac{1}{T-1}\mathrm{tr}\left(M_\iota\right)=\sigma_\epsilon^2.\end{aligned}$$

Although in reality we only observe $\widehat{\tilde{\epsilon}}_{it}$, we can show that the estimation error between $\widehat{\tilde{\epsilon}}_{it}$ and $\tilde{\epsilon}_{it}$ is negligible. Thus by the law of large numbers

$$\widehat{\sigma}_\epsilon^2=\frac{1}{n\left(T-1\right)}\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{\tilde{\epsilon}}_{it}^2\xrightarrow{d}\frac{1}{n\left(T-1\right)}E\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{\epsilon}_{it}^2\right]=\sigma_\epsilon^2$$

is a consistent estimator of the variance. The stated conclusion follows. □

We implicitly assume regularity conditions that allow us to invoke a law of large numbers and a central limit theorem. We ignore those technical details here.

It is important to notice that the within-group demean in FE eliminates all time-invariant explanatory variables, including the intercept. Therefore from FE we cannot obtain the coefficient estimates of these time-invariant variables.

9.2 Random Effect

The random effect estimator pursues efficiency at a knife-edge special case $\mathrm{cov}\left(\alpha_i, x_{it}\right)=0$. As mentioned above, FE is consistent when α_i and x_{it} are uncorrelated. However, an inspection of the covariance matrix reveals that OLS is inefficient.

The starting point is again the original model, while we assume

Assumption RE.1 $E\left[\epsilon_{it}|\alpha_i, \mathbf{x}_i\right] = 0$ and $E\left[\alpha_i|\mathbf{x}_i\right] = 0$.

RE.1 obviously implies $\text{cov}\left(\alpha_i, x_{it}\right) = 0$, so

$$S = \text{var}\left(u_i|\mathbf{x}_i\right) = \sigma_\alpha^2 \mathbf{1}_T \mathbf{1}_T' + \sigma_\epsilon^2 I_T, \text{ for all } i = 1, \ldots, n.$$

Because the covariance matrix is not a scalar multiplication of the identity matrix, OLS is inefficient.

As mentioned before, FE estimation kills all time-invariant regressors. In contrast, RE allows time-invariant explanatory variables. Let us rewrite the original equation as

$$y_{it} = w_{it}\boldsymbol{\beta} + u_{it},$$

where $\boldsymbol{\beta} = \left(\beta_1, \beta_2'\right)'$ and $w_{it} = (1, x_{it})$ are $K+1$ vectors, i.e., $\boldsymbol{\beta}$ is the parameter including the intercept, and w_{it} is the explanatory variables including the constant. The infeasible GLS estimator is

$$\widehat{\boldsymbol{\beta}}^{RE}_{\text{infeasible}} = \left(\sum_{i=1}^{n} \mathbf{w}_i' S^{-1} \mathbf{w}_i\right)^{-1} \sum_{i=1}^{n} \mathbf{w}_i' S^{-1} \mathbf{y}_i = \left(W' \mathbf{S}^{-1} W\right)^{-1} W' \mathbf{S}^{-1} y$$

where $\mathbf{S} = I_T \otimes S$. In practice, σ_α^2 and σ_ϵ^2 in S are unknown, so we seek consistent estimators. Again, we impose a simplifying assumption parallel to FE.2.

Assumption RE.2 $\text{var}\left(\epsilon_i|\mathbf{x}_i, \alpha_i\right) = \sigma_\epsilon^2 I_T$ and $\text{var}\left(\alpha_i|\mathbf{x}_i\right) = \sigma_\alpha^2$.

Under this assumption, we can consistently estimate the variances from the residuals $\widehat{u}_{it} = y_{it} - x_{it}\widehat{\boldsymbol{\beta}}^{RE}$. That is

$$\widehat{\sigma}_u^2 = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \widehat{u}_{it}^2$$
$$\widehat{\sigma}_\epsilon^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T\left(T-1\right)} \sum_{t=1}^{T} \sum_{r=1}^{T} \sum_{r \neq t} \widehat{u}_{it} \widehat{u}_{ir}.$$

Given the estimated variance and covariance, we can construct $\widehat{\mathbf{S}} = \left(\widehat{\sigma}_u^2 - \widehat{\sigma}_\epsilon^2\right) \cdot I_T + \widehat{\sigma}_\epsilon^2 \cdot \mathbf{1}_T \mathbf{1}_T'$ and then follows the feasible GLS (FGLS)

$$\widehat{\boldsymbol{\beta}}^{RE} = \left(W' \widehat{\mathbf{S}}^{-1} W\right)^{-1} W' \widehat{\mathbf{S}}^{-1} y$$

Exercise 9.1. Show that if RE.1 and RE.2 are satisfied, then

$$\left(\widehat{\sigma}_u^2 \left(W' \widehat{\mathbf{S}}^{-1} W\right)^{-1}\right)^{-1/2} \left(\widehat{\boldsymbol{\beta}}^{RE} - \boldsymbol{\beta}_0\right) \xrightarrow{d} N\left(0, I_{K+1}\right).$$

In econometrics practice, the FE estimator is more popular than the RE estimator as the former is consistent in more general conditions.

9.3 Summary

The formula of the FE estimator or the RE estimators is not important because the estimation and inference are automatically handled by econometric packages. What is important is the conceptual difference of FE and RE on their treatment of the unobservable individual heterogeneity.

Panel data is the first generation of economic "big data", as the number of observations of a cross section is multiplied by the number of time periods. It reflected econometrician's pursuit of controlling heterogeneity, so that the OLS estimate is more credible for causal interpretation.

Further reading: Hsiao (2014) is a comprehensive monograph on the topic of panel data. Su et al. (2016) extends fixed effect models to incorporate group heterogeneity.

Zhentao Shi. Nov 8, 2020.

Bibliography

Hsiao, C. (2014). *Analysis of panel data*. Number 54. Cambridge University Press.

Su, L., Z. Shi, and P. C. Phillips (2016). Identifying latent structures in panel data. *Econometrica 84*(6), 2215–2264.