Panel Data

Zhentao Shi

November 3, 2019

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

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

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 theory in econometrics.

1.1 Fixed Effect

OLS is consistent for the linear projection model. Since α_i is unobservable, it is absorbed into the composite error $u_{it} = \alpha_i + \epsilon_{it}$. If $\operatorname{cov}(\alpha_i, x_{it}) = 0$, 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 $\alpha_i, i = 1, \dots, n$. The rest of 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

$$\overline{y}_i = \beta_1 + \overline{x}_i\beta_2 + \bar{u}_{it} = \beta_1 + \overline{x}_i\beta_2 + \alpha_i + \bar{\epsilon}_{it}. \tag{2}$$

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

$$\tilde{y}_{it} = \tilde{x}_{it}\beta_2 + \tilde{\epsilon}_{it} \tag{3}$$

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 $\text{var}(\epsilon_i|\alpha_i, \mathbf{x}_i) = \sigma_\epsilon^2 I_T$.

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 $n(T-1)$, 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

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

We implicitly assume some 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.

1.2 Random Effect

The random effect estimator pursues efficiency at a knife-edge special case $\text{cov}(\alpha_i, x_{it}) = 0$. 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 $E[\epsilon_{it}|\alpha_i, \mathbf{x}_i] = 0$ and $E[\alpha_i|\mathbf{x}_i] = 0$.

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

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

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 $\boldsymbol{\beta} = (\beta_1, \beta_2')'$ 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. Had we known S, the GLS estimator would be

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

where $\mathbf{S} = I_T \otimes S$. ("$\otimes$" denotes the Kronecker product.) 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 $\text{var}(\epsilon_i | \mathbf{x}_i, \alpha_i) = \sigma_\epsilon^2 I_T$ and $\text{var}(\alpha_i | \mathbf{x}_i) = \sigma_\alpha^2$.

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

$$\widehat{\sigma}_u^2 = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \widehat{u}_{it}^2$$

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

Again, we claim the asymptotic normality.

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

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

where $\widehat{\mathbf{S}}$ is a consistent estimator of $\mathbf{S}$.

The complicated formula of the RE estimator is not important because again it will be handled by an econometric package automatically. What is important is the conceptual difference of FE and RE on their treatment of the unobservable individual heterogeneity.