Panel Data Models

Junhui Qian

December 20, 2018

Outline

- Introduction
- The Fixed-Effect Panel Data Model
- The Random-Effect Panel Data Model
- The Dynamic Panel Data Model
- Time Fixed Effects
- Time-Varying Coefficients

Panel Data

- A panel data contain information on the same group of individuals (persons, households, firms, provinces, countries, etc.) over a period of time. See an example of a panel data set next page.
- If a panel data is available, we may deal with endogeneity problems without resorting to IV, at least to some extent.
- Panel data also allow us to control for unobserved factors.
  - Time-invariant factors that differ across individuals or groups.
  - Time effect that are common across individuals or groups.
  - And more generally, factor structure.

An Example of Panel Data

Person	Year	Wage	Gender	Age
1	2001	4000	0	22
1	2002	5000	0	23
1	2003	6000	0	24
2	2001	7000	1	27
2	2002	7500	1	28
2	2003	8000	1	29
3	2001	1500	0	19
3	2002	1600	0	20
3	2003	1650	0	21
		⋮		

A Panel Data Model for Endogeneity Problem

- Suppose that we regress y on x. If some of the elements in x is endogenous, then OLS of $y_i = \beta_0 + x_i'\beta + u_i$ using cross-section data would result in inconsistent estimates. Panel data, with more information on x and y, may help.
- With panel data, under the assumption that u is correlated with x through a time-invariant effect, we may use the following model,

$$y_{it} = x_{it}'\beta + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

where $u_{it} = \mu_i + v_{it}$,

  - μ_i is a time-invariant individual effect, and μ_i may be correlated with x_{it}.
  - For example, in the study of return to education, μ_i can be the unobserved "ability".

Outline

- Introduction
- **The Fixed-Effect Panel Data Model**
- The Random-Effect Panel Data Model
- The Dynamic Panel Data Model
- Time Fixed Effects
- Time-Varying Coefficients

The Fixed-Effect Model

Consider the following panel data model

$$y_{it} = x_{it}'\beta + \mu_i + v_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T.$$

- μ_i is called individual effect, or group effect, controlling for some time-invariant component of y.
- v_{it} is called idiosyncratic error. We generally assume that v_{it} is i.i.d. across i and t, independent of x and μ.
- If we allow μ_i to be correlated with x_{it}, then this model is called the "fixed-effect panel data model".
- If, in contrast, we assume that μ_i is independent from x_{it} and v_{it}, then the model is called the "random-effect panel data model".

Estimating Fixed-Effect Panel Data Model: I

- An obvious approach is to get rid of μ_i by taking first difference of the equation for each individual. Let $\Delta y_{it} \equiv y_{it} - y_{i,t-1}$, we have

$$\Delta y_{it} = \Delta x_{it}' \beta + e_{it},$$

  where $e_{it} = \Delta v_{it}$.
- Now we can estimate β by OLS.
- e_{it} is serially correlated, so OLS would be inefficient.

Estimating Fixed-Effect Panel Data Model: II

- A second approach is to get rid of μ_i by subtracting individual means from each observations. Specifically, let $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and similarly for other variables. In terms of individual means, the model is

  $$\bar{y}_i = \bar{x}_i'\beta + \mu_i + \bar{v}_i.$$

  Subtracting the individual means from the original model, we obtain

  $$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta + (v_{it} - \bar{v}_i).$$

- Now OLS estimates β efficiently.

Estimating Fixed-Effect Panel Data Model: II

It can be shown that the second approach is, in effect, to treat individual effects as coefficients on dummy variables and run least square (LSDV). Specifically, let (y_i, X_i) be the T observations on the i-th individual. We can rewrite our model as

$$y_i = X_i\beta + \iota\mu_i + v_i,\ i = 1, \ldots, N.$$

Or in matrix form,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \beta + \begin{pmatrix} \iota & 0 & \cdots & 0 \\ 0 & \iota & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \iota \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}.$$

Assessing Fixed-Effect Panel Data Model

- Fixed-effects panel data model offers a solution to the endogeneity problem without resorting to IV. Instead, it relies on longer span of data collection on the same individual.
- Fixed-effects model can be consistently estimated as long as the idiosyncratic errors are uncorrelated with the regressors.
- Time-invariant regressors are absorbed by the fixed effects. Thus the effects of time-invariant regressors are unidentified in fixed-effects panel data models. In estimation, it is clear that any time-invariant regressor (e.g., gender, education) would disappear after the first-differencing or de-mean transformation.

Outline

The Random-Effect Panel Data Model

If the individual effects are not correlated with any regressors, i.e., there is no endogeneity problem, then we may use the random-effect panel data model,

$$y_{it} = x_{it}'\beta + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

where $u_{it} = \mu_i + v_{it}$,

- $\mu_i \sim \mathrm{iid}\ \mathrm{N}(0, \sigma_\mu^2)$ is independent from x_{it} and v_{it}
- v_{it} is iid N(0,σ_v^2), independent of x and μ.

The random-effect model can be consistently estimated by OLS, or, more efficiently, GLS.

Estimating Random-Effect Panel Data Model

Note that the covariance matrix of $u = (u_1', \ldots, u_n')'$ has a particular structure,

$$\Omega = \begin{pmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & \Sigma & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \Sigma \end{pmatrix},$$

where

$$\Sigma = \begin{pmatrix} \sigma_\mu^2 + \sigma_v^2 & \sigma_\mu^2 & \cdots & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_v^2 & \cdots & \sigma_\mu^2 \\ & & \vdots & \\ \sigma_\mu^2 & \sigma_\mu^2 & \cdots & \sigma_\mu^2 + \sigma_v^2 \end{pmatrix}$$

Assessing Random-Effect Panel Data Model

- In the random-effect model, time-invariant regressors are no longer absorbed by the fixed effects. Thus the effects of time-invariant regressors are identified in random-effects panel data models.
- When the random-effect assumptions hold, the random-effect approach is more efficient. However, if there is correlation between individual effects and any regressor, then the random-effect approach would yield inconsistent estimation.
- In practice, we use the Hausman-Wu test to check whether the random-effect approach can be employed.

The Hausman-Wu Test

- Since the random-effect estimator $\hat{\beta}_{re}$ is consistent only when μ_i is independent of x and the fixed-effect estimator $\hat{\beta}_{fe}$ is always consistent, we can construct a test statistic based on the distance between $\hat{\beta}_{re}$ and $\hat{\beta}_{fe}$.
- Under the null hypothesis (μ_i is random-effect), we can prove that the covariance matrix $\text{cov}(\hat{\beta}_{re}, \hat{\beta}_{fe} - \hat{\beta}_{fe}) = 0$. Hence we have
$$\Sigma_{\hat{\beta}_{fe} - \hat{\beta}_{re}} = \Sigma_{\hat{\beta}_{fe}} - \Sigma_{\hat{\beta}_{re}}.$$
- Then the Hausman-Wu test statistic is given by
$$W = \left(\hat{\beta}_{fe} - \hat{\beta}_{re}\right)' \hat{\Sigma}^{-1}_{\hat{\beta}_{fe} - \hat{\beta}_{re}} \left(\hat{\beta}_{fe} - \hat{\beta}_{re}\right).$$
Obviously it is in the Wald form. Under the null hypothesis, W has an asymptotic distribution of χ^2_k, where k is the number of elements in β.

The Random-Effects Model with Time-Invariant Regressors

- The random-effect model allows for time-invariant regressors,

$$y_{it} = x_{it}'\beta + z_i'\alpha + \mu_i + v_{it}.$$

  In fixed-effects models, α is not identified.
- But when some elements of x_{it} and z_i are correlated with μ_i, the OLS or GLS estimator for β and α would be inconsistent.
- In this case, we can use instrumental variables (Hausman and Taylor, 1981).

Outline

- Introduction
- The Fixed-Effect Panel Data Model
- The Random-Effect Panel Data Model
- **The Dynamic Panel Data Model**
- Time Fixed Effects
- Time-Varying Coefficients

The Dynamic Panel Data Model

When the right-hand-side variables include the lagged dependent variable, we have a dynamic panel data model:

$$y_{it} = \alpha y_{i,t-1} + x_{it}'\beta + \mu_i + v_{it},$$

where μ_i can be interpreted as fixed or random effects.

- With the lagged dependent variable on the right-hand-side, y_{it} depends on the entire history of x prior to t $(x_{i,t-1}, x_{i,t-2}, \ldots)$ as well as x_{it}, which represents the new information arrived at time t.
- The lagged dependent variable is obviously correlated with μ_i, even when μ_i is uncorrelated with x_{it}.
- And even when we take first difference of the equation, the endogeneity problem remains ($y_{i,t-1}$ is correlated with $v_{i,t-1}$):

$$y_{it} - y_{i,t-1} = \alpha(y_{i,t-1} - y_{i,t-2}) + (x_{it} - x_{i,t-1})'\beta + (v_{it} - v_{i,t-1}).$$

Instrumental Variables for the Dynamic Panel Data Model

Consider the first-difference equation,

$$y_{it} - y_{i,t-1} = \alpha(y_{i,t-1} - y_{i,t-2}) + (x_{it} - x_{i,t-1})'\beta + (v_{it} - v_{i,t-1}),$$

where x is assumed to be exogenous.

- Since $y_{i,t-1}$ is correlated with $v_{i,t-1}$, the above regression suffers from endogeneity.
- Fortunately, there are a lot of ready-to-use instruments. The most obvious are $x_{i,t-1}$ and $y_{i,t-2}$, both of which are correlated with $y_{i,t-1}$ but uncorrelated with $(v_{it} - v_{i,t-1})$.
- And there are other candidates: $x_{i,t-2}$, $x_{i,t-3}$, ..., and $y_{i,t-3}$, $y_{i,t-4}$, ...
- If x_{it} is serially correlated, x_{it} itself may be an instrument. Maybe $x_{i,t+1}$, $x_{i,t+2}$,...? And maybe the "within average" $\bar{x}_i \equiv \frac{1}{T}\sum_{t=1}^{T} x_{it}$.
- However, instruments with distant lags may be weak instruments, which lead to large variance of the IV (GMM) estimator.

The GMM Estimation of the Dynamic Panel Data Model

Rewrite the first-difference equation,

$$\Delta y_{it} = \alpha \Delta y_{i,t-1} + \Delta x_{it}' \beta + \varepsilon_{it},$$

where $\varepsilon_{it} = v_{it} - v_{i,t-1}$. Suppose z_{it} is a vector containing all the IV's, we have $\mathbb{E} z_{it} \varepsilon_{it} = 0$, which is the moment condition. Let

$$\bar{m}(\alpha, \beta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} (\Delta y_{it} - \alpha \Delta y_{i,t-1} - \Delta x_{it}' \beta).$$

The GMM estimator solves

$$\min_{\alpha,\beta} \bar{m}(\alpha, \beta)' W \bar{m}(\alpha, \beta),$$

where W is a positive definite matrix.

The GMM Estimation of the Dynamic Panel Data Model

Consider the original dynamic panel data model,

$$y_{it} = \alpha y_{i,t-1} + x_{it}'\beta + \mu_i + v_{it},$$

Let $u_{it} = \mu_i + v_{it}$, we can also formulate the GMM estimation by looking for appropriate instruments that uncorrelated with $\eta_{it} = u_{it} - \bar{u}_i$.

- Think about it: What instruments would you use?

Outline

A Panel Data Model with Time Fixed Effects

Consider the following panel data model,

$$y_{it} = x_{it}'\beta + \mu_i + \alpha_t + v_{it},$$

where α_t does not change across i.

- α_t is called the time fixed effect. It accounts for the common effect of time-varying factors on all individuals/groups.
- It can be that $\alpha_t = z_t'\gamma$, where z_t is a vector of time series.
- The presence of α_t makes each individual/group have a "time-varying intercept".
- The presence of α_t makes the panel "dependent" across i.

An Example

Suppose that Y_{it} represents the output of firm i at time t and that the firms have the Cobb-Douglas production function,

$$Y_{it} = L_{it}^{\beta_1} K_{it}^{\beta_2} e^{\mu_i} e^{\alpha_t} e^{v_{it}},$$

where L and K represent labor and capital, respectively. Take log on both sides of the equation, we have

$$y_{it} = \beta_1 \ell_{it} + \beta_2 k_{it} + \mu_i + \alpha_t + v_{it},$$

where $y_{it} = \log(Y_{it})$, $\ell_{it} = \log(L_{it})$, $k_{it} = \log(K_{it})$.

- In the above model, μ_i can be interpreted as the firm efficiency.
- α_t may reflect the macro trend in the economy.
- It can be that $\alpha_t = z_t' \gamma$, where z_t include time series such as interest rate, inflation, etc. Of course, γ cannot be identified in this model.

Estimating the Panel Data Model with Time Fixed Effects

Treating both μ_i and α_t as parameters, we can estimate the following

$$y_{it} = x_{it}'\beta + \mu_i + \alpha_t + v_{it}$$

by solving

$$\min_{\{\beta,\{\mu_i\},\{\alpha_t\}\}} \left(y_{it} - x_{it}'\beta - \mu_i - \alpha_t\right)^2 .$$

- When T is fixed and $N \to \infty$, α_t can be consistently estimated.
- When N is fixed and $T \to \infty$, μ_i can be consistently estimated.

The Panel Data Model with Interactive Fixed Effects

Individuals in the panel may react to the common factor differently. To model this, we consider

$$y_{it} = x_{it}'\beta + \lambda_i' f_t + v_{it},$$

where f_t is a $R \times 1$ vector of unobserved common factors and λ_i is the $R \times 1$ vector of factor loadings.

- We assume that the number of factors, R, is known.
- When $\lambda_i = \begin{pmatrix} \mu_i \\ 1 \end{pmatrix}$ and $f_t = \begin{pmatrix} 1 \\ \alpha_t \end{pmatrix}$, the above model reduces to the panel data model with individual and time fixed effects. In this case, $R = 2$.
- Neither λ_i nor f_t can be identified. The factor structure $\lambda_i' f_t$ plays the controlling role.

Estimating the Panel Data Model with Interactive Fixed Effects

Let $Y_t = (y_{1t}, \ldots, yNt)'$, $X_t = (x_{1t}, \ldots, x_{Nt})'$, $\Lambda = (\lambda_1, \ldots, \lambda_N)'$ and $F = (f_1, \ldots, f_T)'$. We may estimate the panel data model with interactive fixed effects by solving

$$\min_{\beta,\Lambda,F} \frac{1}{NT}(Y_t - X_t\beta - \Lambda f_t)'(Y_t - X_t\beta - \Lambda f_t).$$

Concentrating f_t out, the above problem is equivalent to

$$\min_{\beta,\Lambda} \frac{1}{NT}(Y_t - X_t\beta)'(I - P_\Lambda)(Y_t - X_t\beta),$$

where P_Λ is the orthogonal projection on range(Λ).

Estimating the Panel Data Model with Interactive Fixed Effects

The minimization problem

$$\min_{\beta,\Lambda} \frac{1}{NT}(Y_t - X_t\beta)'(I - P_\Lambda)(Y_t - X_t\beta),$$

is further equivalent to

$$\min_{\beta} \frac{1}{N} \sum_{r=R+1}^{N} \mu_r \left(\frac{1}{T} \sum_{t=1}^{T} (Y_t - X_t\beta)(Y_t - X_t\beta)' \right),$$

where $\mu_r(A)$ denotes the r-th largest eigenvalue of A, which can be obtained by principal component analysis.

Determining the Number of Factors

Let

$$V(R, \hat{\beta}_R) = \frac{1}{N} \sum_{r=R+1}^{N} \mu_r \left(\frac{1}{T} \sum_{t=1}^{T} \left(Y_t - X_t \hat{\beta}_R \right) \left(Y_t - X_t \hat{\beta}_R \right)' \right).$$

We can use an information criterion of the following form,

$$IC(R) = \log(V(R, \hat{\beta}_R)) + \rho R,$$

where ρ can choose from

$$\rho = \frac{(N+T)p}{NT} \log \left(\frac{NT}{N+T} \right),$$

or

$$\rho = \frac{(N+T)p}{NT} \log \left(\min \left(\sqrt{N}, \sqrt{T} \right) \right).$$

Outline

- Introduction
- The Fixed-Effect Panel Data Model
- The Random-Effect Panel Data Model
- The Dynamic Panel Data Model
- Time Fixed Effects
- **Time-Varying Coefficients**

Estimating the Panel Data Model with Time-Varying Coefficients

First consider the following panel data model,

$$y_{it} = x_{it}'\beta_t + \mu_i + v_{it},$$

where x_{it} is exogenous (e.g., do not contain the lagged dependent variable) and the coefficient β_t may change over time.

- Even when β_t changes every time, it can be consistently estimated as $N \to \infty$.
- To estimate the model, we can first-difference the above equation,

$$\Delta y_{it} = x_{it}'\beta_t - x_{i,t-1}'\beta_{t-1} + \Delta v_{it}.$$

  Then we obtain the estimator for β_t by solving

$$\min_{\{\beta_t\}} \sum_{i=1}^{N} \sum_{t=2}^{T} \left(\Delta y_{it} - x_{it}'\beta_t + x_{i,t-1}'\beta_{t-1} \right)^2 .$$

When There Are A Few Structural Breaks

Now consider the case where β_t has a few breaks. Let p be the number of breaks. We assume $p << T$.

- Let $\theta_1 = \beta_1$ and $\theta_t = \beta_t - \beta_{t-1}$ for $t \geq 2$. The above statement is equivalent to that $\{\theta_t, t = 2, \ldots, T\}$ is sparse.
- Then we can estimate the model by solving a penalized least square,

$$\min_{\{\theta_t\}} \sum_{i=1}^{N} \sum_{t=2}^{T} \left(\Delta y_{it} - x_{it}'\beta_t + x_{i,t-1}'\beta_{t-1}\right)^2 + \lambda \sum_{t=2}^{T} w_t \|\beta_t - \beta_{t-1}\|,$$

  where λ is a tuning parameter on the group-fused-Lasso penalty and w_t is a weight.
- A natural choice of weight is

$$w_t = \|\tilde{\beta}_t - \tilde{\beta}_{t-1}\|^{-2},$$

  where $\tilde{\beta}_t$ is a preliminary estimate of β_t.

When There Are Endogenous Variables

Now consider the case where x_{it} contains endogenous variables or lagged dependent variable (i.e., dynamic panel).

- In the case where β_t changes at every t, we may estimate the model by solving

$$\min_{\{\beta_t\}} \sum_{t=2}^{T} \left\{ \frac{1}{N} \sum_{i=1}^{N} \rho_{it}\left(\beta_t, \beta_{t-1}\right) \right\}' W_t \left\{ \frac{1}{N} \sum_{i=1}^{N} \rho_{it}\left(\beta_t, \beta_{t-1}\right) \right\},$$

  where $\rho_{it}\left(\beta_t, \beta_{t-1}\right) = z_{it}(\Delta y_{it} - x_{it}'\beta_t + x_{i,t-1}'\beta_{t-1})$, z_{it} is the vector of IV's.

Penalized GMM Estimation

When there may be a few breaks and x_{it} contains endogenous variables or lagged dependent variable (i.e., dynamic panel), we can estimate the model by solving a penalized GMM,

$$\min_{\{\beta_t\}} \sum_{t=2}^{T} \left\{ \frac{1}{N} \sum_{i=1}^{N} \rho_{it}\left(\beta_t, \beta_{t-1}\right) \right\}' W_t \left\{ \frac{1}{N} \sum_{i=1}^{N} \rho_{it}\left(\beta_t, \beta_{t-1}\right) \right\} + \lambda \sum_{t=2}^{T} w_t \|\beta_t - \beta_{t-1}\|,$$

where $\rho_{it}\left(\beta_t, \beta_{t-1}\right) = z_{it}(\Delta y_{it} - x_{it}'\beta_t + x_{i,t-1}'\beta_{t-1})$, z_{it} is the vector of IV's, and w_t is a weight.

Time-Varying Coefficient and Interactive Fixed Effects

Finally we consider the time-varying-coefficient model with interactive fixed effects,

$$y_{it} = x_{it}'\beta_t + \lambda_i' f_t + v_{it},$$

where f_t is a $R \times 1$ vector of unobserved common factors and λ_i is the $R \times 1$ vector of factor loadings.

- We can estimate the model by solving

$$\min_{\{\beta_t\},\Lambda,F} \frac{1}{NT}(Y_t - X_t\beta_t - \Lambda f_t)'(Y_t - X_t\beta_t - \Lambda f_t).$$

- By concentrating Λ and F out, the above is equivalent to

$$\min_{\beta} \frac{1}{N} \sum_{r=R+1}^{N} \mu_r \left(\frac{1}{T} \sum_{t=1}^{T} (Y_t - X_t\beta_t)(Y_t - X_t\beta_t)' \right),$$

  where $\mu_r(A)$ denotes the r-th largest eigenvalue of A, which can be obtained by principal component analysis.

Penalized Principal Component (PPC) Estimation

When there may be a few structural breaks in β_t, we can estimate the model by solving

$$\min_{\{\beta_t\},\Lambda,F} \frac{1}{NT}(Y_t - X_t\beta_t - \Lambda f_t)'(Y_t - X_t\beta_t - \Lambda f_t) + \gamma \sum_{t=2}^{T} w_t \|\beta_t - \beta_{t-1}\|.$$

where γ is a tuning parameter on the group-fused-Lasso penalty and w_t is a weight. The above problem is equivalent to

$$\min_{\beta} \frac{1}{N} \sum_{r=R+1}^{N} \mu_r \left(\frac{1}{T} \sum_{t=1}^{T} (Y_t - X_t\beta_t)(Y_t - X_t\beta_t)' \right) + \gamma \sum_{t=2}^{T} w_t \|\beta_t - \beta_{t-1}\|.$$

Since we minimize the component using principal component analysis, we call this procedure the penalized principal component (PPC) estimation.

References

- Qian, J., L. Su, 2016, Shrinkage estimation of common breaks in panel data models via adaptive group fused Lasso, Journal of Econometrics 191, 86-109.
- Li, D., J. Qian, L. Su, 2016, Panel data models with interactive fixed effects and multiple structural breaks, Journal of the American Statistical Association, 111, 1804-1819.

Concluding Remarks

- Panel data allow us to control for unobserved factors, thus making us more resourceful in dealing with the endogeneity problem.
- The present lecture has not exhausted the possibilities of modeling panel data.
  - The panel data model can be heterogenous, in that coefficients for each individual can be different.
  - The heterogeneity can have some group structure, with possibly structural breaks.
  - The individual/group effect can be non-additive.