Practical notes on panel data models with interactive effects

Bai, Jushan and Li, Kunpeng

Columbia University, Capital University of Economics and Business

1 September 2017

PRACTICAL NOTES ON PANEL DATA MODELS WITH INTERACTIVE EFFECTS

JUSHAN BAI AND KUNPENG LI

ABSTRACT. This note is intended for researchers who want to use the interactive effects model for empirical modeling. We consider how to estimate interactive effects models when some of the factors and factor loading are observable. Observable factors are common regressors which do not vary across individuals such as macroeconomic variables, but their regression coefficients are individual-dependent. Observable factor loadings correspond to time-invariant regressors such that race, gender and education, but their regression coefficients are time dependent. This note elaborates the estimation procedures in Bai (2009) in the presence of such regressors.

Keywords: observable factors, observable factor loadings, common regressors, time-invariant regressors

1. OBSERVABLE FACTORS

Consider the model

$$y_{it} = x_{it}'\beta + \phi_i' g_t + \lambda_i' f_t + u_{it}$$

with

$$i = 1, 2, ..., N; t = 1, ..., T.$$

Observable variables are (y_{it}, x_{it}, g_t), all the rest are unobservable. The regressors g_t are observable, but they do not vary over i. We refer g_t as the common regressors. These can be policy variables or macroeconomic variables (e.g., interest rates, unemployment, inflation etc). Write the model in vector form

$$Y_i = X_i\beta + G\phi_i + F\lambda_i + u_i,$$

where Y_i is $T \times 1$, $X_i = (x_{i1}, x_{i2}, .., x_{iT})'$, $G = (g_1, g_2, ..., g_T)'$, and $F = (f_1, f_2, ..., f_T)'$. Define the projection matrix

$$M_G = I_T - G(G'G)^{-1}G'.$$

Do transformation and use $M_G G = 0$,

$$M_G Y_i = M_G X_i \beta + M_G F \lambda_i + M_G u_i.$$

Renaming variables

$$Y_i^* = X_i^* \beta + F^* \lambda_i + u_i^*$$

where $Y_i^* = M_G Y_i$, etc. The transformation eliminates the observable factors. Data in "*" form conforms with the model in Bai (2009). From here, we can use the method in Bai (2009) to estimate β, F^*, and λ_i for each i. A MATLAB program is available for estimating the model (https://ideas.repec.org/c/boc/bocode/m430011.html).

Jushan Bai: Department of Economics, Columbia University.
Kunpeng Li: International School of Economics and Management, Capital University of Economics and Business, Beijing, China.

Next to estimate ϕ_i, the coefficients of the common regressors g_t, we have to make an assumption that G and F are orthogonal, that is, $G'F = 0$. Otherwise, ϕ_i is not identifiable (ϕ_i and λ_i are not separable). Orthogonality between G and F implies $F^* = M_G F = F$. Now given that β, F and λ_i are estimable, moving relevant terms to the left hand side

$$Y_i - X_i\beta - F\lambda_i = G\phi_i + u_i$$

or

$$Y_i^\dagger = G\phi_i + u_i$$

where $Y_i^\dagger = Y_i - X_i\beta - F\lambda_i$, which can be assumed known (at least it is estimable). Least squares regression of $Y_i^\dagger$ on G gives an estimate for ϕ_i. That is,

$$\hat{\phi}_i = (G'G)^{-1}G^{-1}Y_i^\dagger, \quad i = 1, 2, ..., N$$

2. OBSERVABLE FACTOR LOADINGS

Suppose there are regressors that are time-invariant (equivalent to factor loadings being observable)

$$y_{it} = x_{it}'\beta + z_i'\gamma_t + \lambda_i' f_t + u_{it}$$

Observable variables are (y_{it}, x_{it}, z_i), where z_i is time invariant, such as race, gender and education. Write the model as

$$Y_t = X_t\beta + Z\gamma_t + \Lambda f_t + u_i$$

where $Y_t = (y_{1t}, y_{2t}, ..., y_{Nt})'$ is $N \times 1$; $Z = (z_1, ..., z_N)'$, and $\Lambda = (\lambda_1, ..., \lambda_N)'$. Let $M_Z = I_N - Z(Z'Z)^{-1}Z'$, and do transformation

$$M_Z Y_t = M_Z X_t\beta + M_Z\Lambda f_t + M_Z u_i$$

or more compactly

$$\dot{Y}_t = \dot{X}_t\beta + \dot{\Lambda} f_t + \dot{u}_i$$

where $\dot{Y}_t = M_Z Y_t$, etc. The model again conforms with that of Bai (2009) so that β, $\dot{\Lambda}$, f_t are estimable. To estimate γ_t, we assume Z and Λ are orthogonal (i.e., $Z'\Lambda = 0$) to eliminate rotational indeterminacy. Then $\dot{\Lambda}$ coincides with Λ. Notice

$$Y_t - X_t\beta - \Lambda f_t = Z\gamma_t + u_t$$

or

$$\ddot{Y}_t = Z\gamma_t + u_t$$

with $\ddot{Y}_t = Y_t - X_t\beta - \dot{\Lambda} f_t$. Least squares regression of $\ddot{Y}_t$ on Z gives the estimator of γ_t:

$$\hat{\gamma}_t = (Z'Z)^{-1}Z'\ddot{Y}_t, \quad t = 1, 2..., T$$

3. OBSERVABLE FACTOR AND FACTOR LOADINGS

A more general model is the presence of both common regressors g_t and time-invariant regressors z_i:

$$y_{it} = x_{it}'\beta + z_i'\gamma_t + \phi_i' g_t + \lambda_i' f_t + u_{it}$$

Observable variables are $(y_{it}, x_{it}, g_t, z_i)$. For this case, see Bai and Li (2014) for details.

Suppose that x_{it} is a $k \times 1$ vector (containing k regressors) such that $x_{it} = (x_{it,1}, ..., x_{it,k})'$. Define the $N \times T$ matrix $X_j = [x_{it,j}]_{N\times T}$ for the jth explanatory variable ($j = 1, 2, ..., k$). Write the model as

$$Y = X_1\beta_1 + \cdots + X_k\beta_k + Z\Gamma' + \Phi G' + \Lambda F' + u \tag{3.1}$$

so each term is an $N \times T$ matrix, then left multiply M_Z and right multiply M_G to get

$$M_Z Y M_G = (M_Z X_1 M_G)\beta_1 + ... + (M_Z X_k M_G)\beta_k + (M_Z\Lambda)(F'M_G) + M_Z u M_G \tag{3.2}$$

or equivalently

$$\ddot{Y} = \ddot{X}_1\beta_1 + \cdots + \ddot{X}_k\beta_k + \ddot{\Lambda}\ddot{F}' + \ddot{u}$$

with $\ddot{Y} = M_Z Y M_G$, $\ddot{X}_j = M_Z X_j M_G$, $\ddot{u} = M_Z u M_G$, $\ddot{\Lambda} = M_Z\Lambda$ and $\ddot{F} = M_G F$. The transformation eliminates both the observable and time-invariant regressors. The model now reduces to that of Bai (2009) again.

Let $\hat{\beta}, \widehat{\ddot{\Lambda}}$ and $\widehat{\ddot{F}}$ be the estimator of Bai (2009) for the above model and define the residuals with the original data (untransformed) Y and X_j

$$\widetilde{Y} = Y - X_1\hat{\beta}_1 - \cdots - X_k\hat{\beta}_k - \widehat{\ddot{\Lambda}}\widehat{\ddot{F}}'.$$

Under the identification condition $Z\perp\Lambda$, $Z\perp\Phi$, $F\perp G$, we estimate Φ, the coefficient matrix of common regressors, by

$$\widehat{\Phi} = (M_Z\widetilde{Y}G)(G'G)^{-1}$$

and estimate Γ, the coefficient matrix of time invariant regressors, by

$$\widehat{\Gamma} = (\widetilde{Y} - \widehat{\Phi}G')'Z(Z'Z)^{-1}$$

Bai and Li (2014) also study the maximum likelihood estimation of the model.

Remarks: One can equally work with $T \times N$ data matrices, where each column represents a time series, and ith column belongs to the ith individual. That will simply be the transpose of the above model. Alternatively one can stack the data into a long vector. From $\text{vec}(ABC) = (C' \otimes A)\text{vec}(B)$, vectorization of (3.1) gives

$$\mathbf{y} = \mathbf{x}\beta + (I_T \otimes Z)\text{vec}(\Gamma') + (G \otimes I_N)\text{vec}(\Phi) + \text{vec}(\Lambda F') + \text{vec}(u)$$

where $\mathbf{y} = \text{vec}(Y)$, $\mathbf{x} = [\text{vec}(X_1), ..., \text{vec}(X_k)]_{NT\times k}$. Left multiplying the matrix $M_G \otimes M_Z$ will eliminate terms involving Z and G. This will be equivalent to vectoring (3.2). The idea is that whether one works with data matrices of format $N \times T$ or $T \times N$ or long format $NT \times k$, transformation can be easily performed. Notationwise, equation (3.1) or its transpose appears to be easier.

The preceding discussion assumes the coefficients of z_i are time varying, and the coefficients of g_t are individual-dependent. Now consider the model in which these coefficients are constant

$$y_{it} = x_{it}'\beta + z_i'\gamma + g_t'\phi + \lambda_i'f_t + u_{it}.$$

Let $\widetilde{x}_{it} = (x_{it}', z_i', g_t')'$ and $\theta = (\beta', \gamma', \phi')'$, the above model is equivalent to

$$y_{it} = \widetilde{x}_{it}'\theta + \lambda_i'f_t + u_{it}.$$

We can still use the estimation procedure of Bai (2009) to estimate θ. It is important to note that we cannot further allow unobservable additive fixed effects in this model. That is, the interactive effects $\lambda_i'f_t$ must be genuine. Mathematically, this requires $[1_N, Z, \Lambda]$ be of a full column rank, where 1_N is an $N \times 1$ vector of

1s. Also, $[1_T, G, F]$ must be of full column rank (see Bai, 2009). The full-column rank assumptions are necessary for identification of γ and ϕ. These assumptions permit z_i to be correlated with the unobservable λ_i and g_t to be correlated with the unobservable f_t, an important feature of panel data models.

References

Bai, J. (2009): Panel data models with interactive fixed effects, *Econometrica*, 77 (4), 1229-1279.

Bai, J. and K.P. Li (2014): Theory and methods of panel data models with interactive effects *The Annals of Statistics* 42 (1), 142-170