EC708 Discussion 7
Linear Panel and Numerical Optimization

Yan Liu

Department of Economics

Boston University

March 24, 2023

Outline

1. Linear Panel
    - Relationship between FE and FD Estimators
    - Arellano-Bond: Weak Instruments
2. Numerical Optimization
    - Full-Newton Methods
    - Quasi-Newton Methods

Table of Contents

Linear Panel

Relationship between FE and FD Estimators

Static Linear Panel Data Model:

$$y_{it} = x_{it}'\beta + \alpha_i + u_{it}.$$

Take first-differences to eliminate α_i:

$$\underbrace{y_{it} - y_{it-1}}_{\Delta y_{it}} = (\underbrace{x_{it} - x_{it-1}}_{\Delta x_{it}})'\beta + \underbrace{u_{it} - u_{it-1}}_{\Delta u_{it}}.$$

A compact form:

$$DY_i = DX_i\beta + Du_i, \text{ where } \underset{(T-1)\times T}{D} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}.$$

Linear Panel

Relationship between FE and FD Estimators

FD estimator is OLS applied to $DY_i = DX_i\beta + Du_i$:

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

- When $T = 2$,

$$D'D = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = 2Q_T \Rightarrow \hat{\beta}_{FD} = \hat{\beta}_{FE}.$$

- When $T > 2$, what is the relationship between $\hat{\beta}_{FD}$ and $\hat{\beta}_{FE}$?

Linear Panel

Relationship between FE and FD Estimators

Assumptions on u_{it}:

1. (strict exogeneity): $E[u_{it}|X_i, \alpha_i] = 0$.
2. (homoskedasticity and no serial correlation) $E[u_i u_i'|X_i, \alpha_i] = \sigma_u^2 I_T$.

Then,

- $\hat{\beta}_{FD}$ is unbiased and consistent: $E[Du_i|X_i] = 0$.
- $\hat{\beta}_{FD}$ is not efficient: $E[(Du_i)(Du_i)'|X_i] = \sigma_u^2 DD'$ is not spherical.

Linear Panel

Relationship between FE and FD Estimators

The optimal estimator is given by GLS:

$$\hat{\beta}_{FD,GLS} = \left(\sum_{i=1}^{N} X_i' D' (DD')^{-1} D X_i \right)^{-1} \sum_{i=1}^{N} X_i' D' (DD')^{-1} D Y_i.$$

What is $D'(DD')^{-1}D$? Trick: let

$$\underset{T \times T}{H} = \begin{bmatrix} T^{-1/2} \mathbf{1}_T' \\ (DD')^{-1/2} D \end{bmatrix}.$$

Linear Panel

Relationship between FE and FD Estimators

We can verify that $D\mathbf{1}_T = \underset{(T-1)\times 1}{\mathbf{0}}$ and so

$$HH' = \begin{bmatrix} T^{-1}\mathbf{1}_T'\mathbf{1}_T & T^{-1}\mathbf{1}_T'D'(DD')^{-1/2} \\ T^{-1}(DD')^{-1/2}D\mathbf{1}_T & (DD')^{-1/2}DD'(DD')^{-1/2} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \underset{1\times(T-1)}{\mathbf{0}} \\ \underset{(T-1)\times 1}{\mathbf{0}} & I_{T-1} \end{bmatrix} = I_T.$$

Hence,

$$I_T = (HH')' = H'H = T^{-1}\mathbf{1}_T\mathbf{1}_T' + D'(DD')^{-1}D$$

$$\Rightarrow D'(DD')^{-1}D = Q_T \text{ and } \hat{\beta}_{FD,GLS} = \hat{\beta}_{FE}!$$

Remarks.

- Alternatively, if $E[(Du_i)(Du_i)'|X_i, \alpha_i] = \sigma_e^2 I_{T-1}$, $\hat{\beta}_{FD}$ is efficient.
  - Now u_{it} is a random walk: $E[u_{it}u_{is}] = s\sigma_e^2$ for $t > s$.
- What does the GLS transformation do?

$$(DD')^{-1/2}DY_i = (DD')^{-1/2}DX_i\beta + (DD')^{-1/2}Du_i.$$

It turns out that each row of $(DD')^{-1/2}DY_i$ is

$$\sqrt{\frac{T-t}{T-t+1}}\left[y_{it} - \frac{1}{T-t}(y_{it+1} + \cdots + y_{iT})\right].$$

$\Rightarrow$ "Forward orthogonal transformation"

Table of Contents

Linear Panel

Arellano-Bond: Weak Instruments

Dynamic Linear Panel Data Model:

$$y_{it} = y_{it-1}\beta + \alpha_i + u_{it}.$$

Consider the first-differenced equation:

$$\Delta y_{it} = \beta \Delta y_{it-1} + \Delta u_{it}.$$

Assumption 4.6 (Sequential Exogeneity). $E[u_{it}|y_{it-1}, \ldots, y_{i1}, \alpha_i] = 0.$

- At $t > 2$, $(y_{i1}, \ldots, y_{it-2})$ are valid IVs for $\Delta y_{it-1} = y_{it-1} - y_{it-2}$.
- $\frac{(T-1)(T-2)}{2}$ moment conditions:

$$E[(\Delta y_{it} - \beta \Delta y_{it-1}) y_{is}] = 0, \quad t = 3, \ldots, T, \ s = 1, \ldots, t-2.$$

Linear Panel

Arellano-Bond: Weak Instruments

For simplicity, consider $T = 3$:

$$y_{i2} = y_{i1}\beta + \alpha_i + u_{i2}.$$

Subtract y_{i1} from both sides:

$$\Delta y_{i2} = (\beta - 1)y_{i1} + \alpha_i + u_{i2}.$$

Under sequential exogeneity, the first-stage coefficient on instrument y_{i1} is

$$\pi = \frac{E[y_{i1}\Delta y_{i2}]}{E[y_{i1}^2]} = (\beta - 1) + \frac{E[y_{i1}\alpha_i]}{E[y_{i1}^2]}$$

Linear Panel

Arellano-Bond: Weak Instruments

Assuming stationarity ($|\beta| < 1$), backwards recursion yields

$$y_{it} = \sum_{s=0}^{\infty} \beta^s (\alpha_i + u_{it-s}) = \frac{\alpha_i}{1-\beta} + \sum_{s=0}^{\infty} \beta^s u_{it-s}.$$

Hence,

$$E[y_{i1}\alpha_i] = E\Big[\Big(\frac{\alpha_i}{1-\beta} + \sum_{s=0}^{\infty} \beta^s u_{it-s}\Big)\alpha_i\Big] = \frac{\sigma_\alpha^2}{1-\beta},$$

$$E[y_{i1}^2] = E\Big[\Big(\frac{\alpha_i}{1-\beta} + \sum_{s=0}^{\infty} \beta^s u_{it-s}\Big)^2\Big] = \frac{\sigma_\alpha^2}{(1-\beta)^2} + \frac{\sigma_u^2}{1-\beta^2}.$$

Some algebra shows

$$\pi = (\beta - 1)\frac{k}{k + \sigma_\alpha^2/\sigma_u^2}, \quad k = \frac{1-\beta}{1+\beta}.$$

Linear Panel

Arellano-Bond: Weak Instruments

The Arellano-Bond estimator suffers from weak instruments if

- β is close to 1 (near unit root);
- $\sigma_\alpha^2/\sigma_u^2$ is large (fixed effect dominates idiosyncratic effect).

To reduce the weak instrument problem, Blundell and Bond (1998) proposed to introduce additional moment conditions. Recall

$$y_{it} = y_{it-1}\beta + \alpha_i + u_{it}.$$

Blundell-Bond instrument: Δy_{it-1} is an instrument for y_{it-1} if

- Δy_{it-1} is correlated with y_{it-1};
- Δy_{it-1} is uncorrelated with $\alpha_i + u_{it} \Rightarrow$ need $E[\Delta y_{it-1}\alpha_i] = 0$.

Linear Panel

Arellano-Bond: Weak Instruments

A sufficient condition for $E[\Delta y_{it-1}\alpha_i] = 0$ is

$$E\left[\left(y_{i1} - \frac{\alpha_i}{1-\beta}\right)\alpha_i\right] = 0.$$

Why? Applying backwards recursion to $\Delta y_{it-1} = \beta \Delta y_{it-2} + \Delta u_{it-1}$ yields

$$\Delta y_{it-1} = \beta^{t-3}\Delta y_{i2} + \sum_{s=0}^{t-3} \beta^s \Delta u_{it-1-s},$$

where

$$\Delta y_{i2} = (\beta - 1)y_{i1} + \alpha_i + u_{i2} = (\beta - 1)\left(y_{i1} - \frac{\alpha_i}{1-\beta}\right) + u_{i2}.$$

Put together,

$$E[\Delta y_{it-1}\alpha_i] = \beta^{t-3}(\beta - 1)E\left[\left(y_{i1} - \frac{\alpha_i}{1-\beta}\right)\alpha_i\right].$$

Linear Panel

Arellano-Bond: Weak Instruments

Blundell and Bond (1998) proposed combining Arellano-Bond moments with the level moments:

$$E[(y_{it} - y_{it-1}\beta)\Delta y_{it-1}] = 0, \quad t = 2, \ldots, T.$$

Implemented in Stata (command xtdpdsys or xtdpd).

Table of Contents

Numerical Optimization

Full-Newton Methods: Newton-Raphson

Goal: find the MLE $\hat{\theta}_T \in \arg\max_{\theta \in \Theta} \bar{\ell}_T(\theta)$.

General algorithm:

- Specify starting value θ_0
- In each iteration, move to a new parameter value at which $\bar{\ell}_T(\theta)$ is higher than at the previous parameter value

Given current value θ_k, what is the best value for θ_{k+1}?

Numerical Optimization

Full-Newton Methods: Newton-Raphson

Take a second-order Taylor's approximation of $\bar{\ell}_T(\theta_{k+1})$ around $\bar{\ell}_T(\theta_k)$:

$$\bar{\ell}_T(\theta_{k+1}) \approx \bar{\ell}_T(\theta_k) + (\theta_{k+1} - \theta_k)' s_k + \frac{1}{2}(\theta_{k+1} - \theta_k)' H_k (\theta_{k+1} - \theta_k).$$

Find θ_{k+1} that maximizes this approximation:

$$s_k + H_k(\theta_{k+1} - \theta_k) = 0 \;\Rightarrow\; \theta_{k+1} = \theta_k - \underbrace{H_k^{-1}}_{\text{step size}} \cdot \underbrace{s_k}_{\text{direction}}.$$

Iterate until convergence, which can be defined in many ways:

- $\bar{\ell}_T(\theta_{k+1})$ close to $\bar{\ell}_T(\theta_k)$
- θ_{k+1} close to θ_k
- s_{k+1} close to s_k

Numerical Optimization

Full-Newton Methods: Newton-Raphson

If $\bar{\ell}_T(\theta)$ were exactly quadratic, then Newton-Raphson reaches the maximum in one step from any starting value. Consider θ being scalar:

$$\bar{\ell}_T(\theta) = a + b\theta + c\theta^2.$$

The maximum is $\hat{\theta}_T = -\frac{b}{2c}$. The gradient and Hessian are

$$s_k = b + 2c\theta_k, \quad H_k = 2c.$$

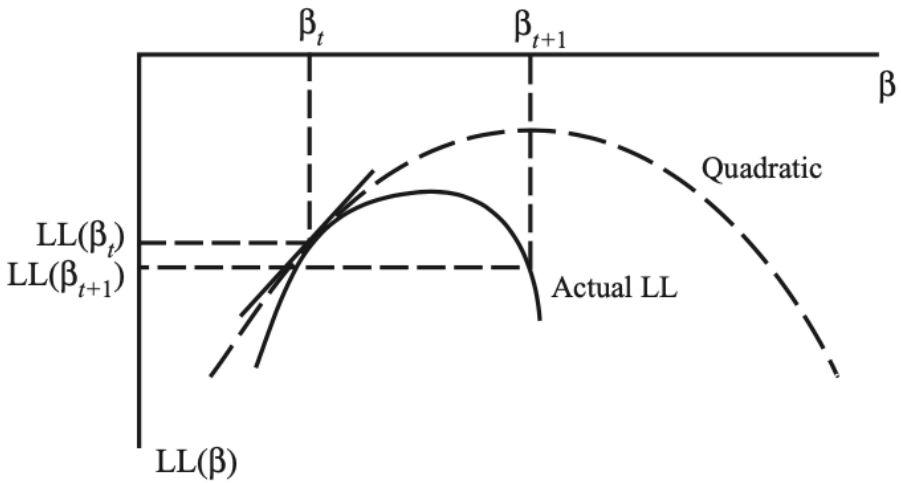
Hence, Newton-Raphson gives

$$\begin{aligned}\theta_{k+1} &= \theta_k - H_k^{-1} s_k \\ &= \theta_k - \frac{1}{2c}(b + 2c\theta_k) \\ &= -\frac{b}{2c} = \hat{\theta}_T.\end{aligned}$$

Numerical Optimization

Full-Newton Methods: Newton-Raphson

Most log-likelihood functions are not quadratic.

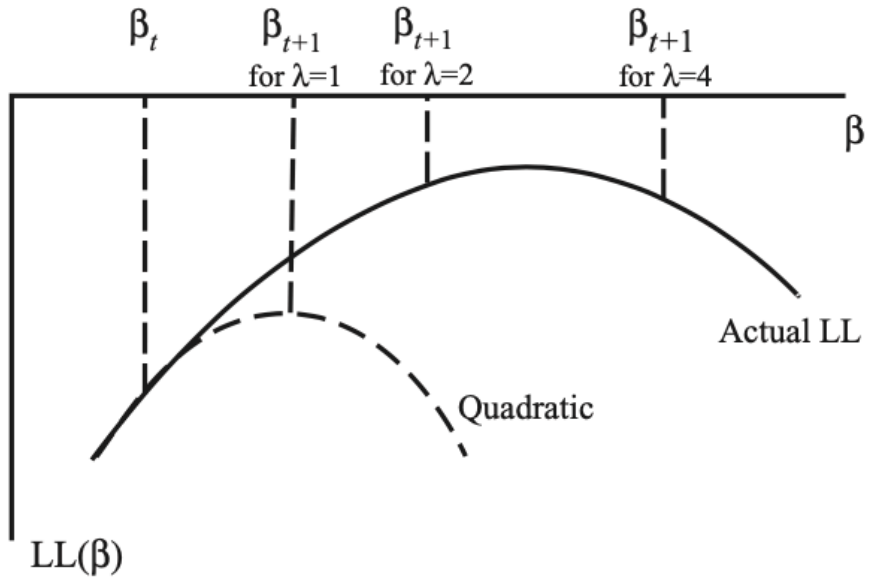


- Newton-Raphson may step past $\hat{\theta}_T$ and move to a lower $\bar{\ell}_T(\theta)$.
- Introduce a **step size** λ_k to ensure each iteration provides an increase in $\bar{\ell}_T(\theta)$:

$$\theta_{k+1} = \theta_k - \lambda_k H_k^{-1} s_k.$$

Numerical Optimization

Full-Newton Methods: Newton-Raphson



Step-size adjustment: Start with $\lambda_k = 1$.

- If $\bar{\ell}_T(\theta_{k+1}) < \bar{\ell}_T(\theta_k)$, continue halving λ_k until $\bar{\ell}_T(\theta_{k+1}) > \bar{\ell}_T(\theta_k)$.
- If $\bar{\ell}_T(\theta_{k+1}) > \bar{\ell}_T(\theta_k)$, continue doubling λ_k as long as doing so further raises $\bar{\ell}_T(\theta_{k+1})$.

Numerical Optimization

Full-Newton Methods: Newton-Raphson

Drawbacks:

- Calculation of the Hessian is usually computation-intensive.
  - $\frac{K(K+1)}{2}$ functions to evaluate in each iteration
  - Numerically calculated Hessian might be ill-behaved (singular).
- If the log-likelihood function is not globally concave, no guarantee of an increase in each iteration.
  - Hessian may not be negative definite.
  - Remedy: regularization. Instead of H_k^{-1}, use

$$(H_k + \mu_k I_K)^{-1}, \text{ where } \mu_k < 0.$$

Numerical Optimization

Full-Newton Methods: Gauss-Newton

Consider the model

$$Y_t = f(X_t; \theta) + U_t, \quad U_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), U_t \perp \mathbf{X}.$$

Assume σ^2 is known. The (conditional) log-likelihood function is

$$\bar{\ell}_T(\theta) = -\frac{T}{2} \ln 2\pi - \frac{T}{2} \ln \sigma^2 - \sum_{t=1}^{T} (Y_t - f(X_t; \theta))^2.$$

Numerical Optimization

Full-Newton Methods: Gauss-Newton

The score and Hessian are

$$s(\theta) = 2\sum_{t=1}^{T} \frac{\partial f}{\partial \theta} U_t, \quad H(\theta) = 2\sum_{t=1}^{T} \left(\frac{\partial^2 f}{\partial \theta \partial \theta'} U_t - \frac{\partial f}{\partial \theta} \frac{\partial f}{\partial \theta'} \right).$$

Gauss-Newton iteration at $\theta = \theta_k$:

$$\theta_{k+1} = \theta_k - \left(\sum_{t=1}^{T} - \frac{\partial f}{\partial \theta} \frac{\partial f}{\partial \theta'} \right)^{-1} \Big|_{\theta=\theta_k} \sum_{t=1}^{T} \frac{\partial f}{\partial \theta} U_t \Big|_{\theta=\theta_k}.$$

- Does not require second-order derivatives
- Has an OLS interpretation.

Table of Contents

Numerical Optimization

Quasi-Newton Methods: Overview

Quasi-Newton update takes the form

$$\theta_{k+1} = \theta_k - {G_k}^{-1} s_k$$

for some approximation G_k of H_k. We want

- G_k is easy to compute;
- linear system $G_k \Delta = -s_k$ is easy to solve.

Numerical Optimization

Quasi-Newton Methods: Berndt-Hall-Hall-Hausman (BHHH)

Berndt et al. (1974) utilized the fact that the objective function is the sum of log-likelihoods and proposed to use scores to approximate Hessian. Define

$$G_k = -\frac{1}{T}\sum_{t=1}^{T} s_t(\theta_k)s_t(\theta_k)'.$$

Why does BHHH work? Recall the information matrix equality:

$$E[s_t(\theta_0)s_t(\theta_0)'] = -H_0.$$

Drawbacks: BHHH can give small steps when far from the maximum.

Numerical Optimization

Quasi-Newton Methods: BFGS and DFP

- BHHH obtains G_k solely based on s_k in each iteration.
- BFGS and DFP updates G_k from the previous iteration.

Given θ_{k+1}, θ_k, s_{k+1}, s_k, and G_k, what are the requirements for G_{k+1}?

- $s_{k+1} = s_k + G_{k+1}(\theta_{k+1} - \theta_k)$ (secant method);
- G_{k+1} is symmetric and negative definite.

To ease notation, let $\Delta_k = \theta_{k+1} - \theta_k$ and $\gamma_k = s_{k+1} - s_k$.

Numerical Optimization

Quasi-Newton Methods: Broyden-Fletcher-Goldfarb-Shanno (BFGS)

BFGS uses a rank-two matrix update:

$$G_{k+1} = G_k + auu' + bvv'.$$

Multiply Δ_k on both sides:

$$\underbrace{G_{k+1}\Delta_k}_{=s_{k+1}-s_k=\gamma_k} - G_k\Delta_k = (au'\Delta_k)u + (bv'\Delta_k)v.$$

Putting $u = \gamma_k$, $v = G_k\Delta_k$, and solving for a, b, we get

$$G_{k+1} = G_k - \frac{G_k\Delta_k\Delta_k'G_k}{\Delta_k'G_k\Delta_k} + \frac{\gamma_k\gamma_k'}{\gamma_k'\Delta_k}.$$

BFGS is the algorithm behind Matlab's fminunc.

Numerical Optimization

Quasi-Newton Methods: Davidon-Fletcher-Powell (DFP)

DFP updates G_{k+1}^{-1} instead:

$$G_{k+1}^{-1} = G_k^{-1} + auu' + bvv'.$$

Multiply γ_k on both sides:

$$\underbrace{G_{k+1}^{-1}\gamma_k}_{=\Delta_k} - G_k^{-1}\gamma_k = (au'\gamma_k)u + (bv'\gamma_k)v.$$

Putting $u = \Delta_k$, $v = G_k^{-1}\gamma_k$, and solving for a, b, we get

$$G_{k+1}^{-1} = G_k^{-1} - \frac{G_k^{-1}\gamma_k\gamma_k' G_k^{-1}}{\gamma_k' G_k^{-1}\gamma_k} + \frac{\Delta_k\Delta_k'}{\Delta_k'\gamma_k}.$$

DFP is not as popular as BFGS. There is some evidence that BFGS is more efficient than DFP.