PANEL DATA

(M2 ETE Econometrics II)

Koen Jochmans

Toulouse School of Economics

Last revised on March 10, 2024

Outline

Structure of panel data

Panel (or longitudinal) data contain multiple data points on the same units.

Units can be individuals, firms, or countries, for example.

Units are often followed over time

Examples are

- labor income and hours worked of individuals by week;
- profit and output of firms by quarter;
- gdp and inflation of countries by year.

Other types of repeated measurements are

- student test scores on multiple subjects (at a given time);
- test scores by different students of the same classroom;
- birth weight of children born to the same mother.

Panel data has a multi-index structure, indicating both the unit and time.

We write (y_{it}, x_{it}) for units $i = 1, \ldots, N$ at time $t = 1, \ldots, T$.

A micro-panel has $N >> T$.

Usually enough to look at a two-wave panel.

Here our focus is on controlling for heterogeneity across units.

A macro-panel has $N << T$.

More like VAR-type modeling. Less attention to this in this module.

An example

Suppose that we have

$$y_{it} = x_{it}'\beta + u_{it}, \qquad u_{it} = \alpha_i + \varepsilon_{it}, \qquad \mathbb{E}(\varepsilon_{it}|x_{i1}, \ldots, x_{iT}, \alpha_i) = 0,$$

Say an agricultural (log-linearized) Cobb-Douglas production function.

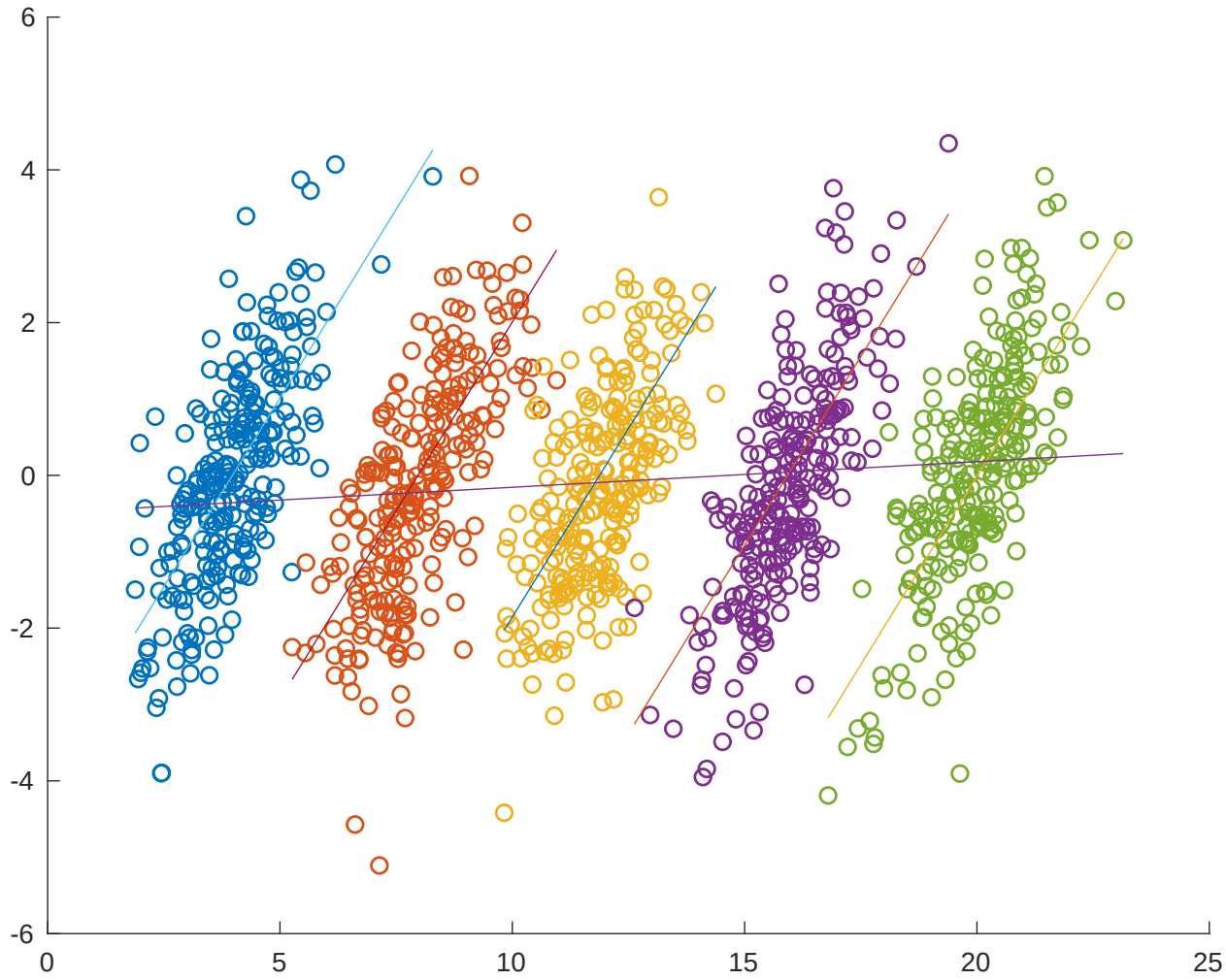
- y_{it} is output;
- x_{it} are observable inputs ;
- α_i is soil quality;
- ε_{it} is rainfall (unpredictable).

Farmer observes (α_i, x_{it}). In general, the inputs x_i, α_i are not uncorrelated. The problem is that α_i is not observed in data. (Otherwise, could just include it in x_i.)

Estimating

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

by (pooled) least-squares suffers from endogeneity bias.


6
4
2
0
-2
-4
-6
0
5
10
15
20
25

There is no information on β in the data in levels.

Information in first differences is useful:

$$\Delta y_{it} := y_{it} - y_{it-1} = \Delta x_{it}'\beta + \Delta\varepsilon_{it},$$

as these do not depend on the α_i.

Furthermore, $\mathbb{E}(\Delta\varepsilon_{it}|x_{i1},\ldots,x_{iT}) = 0$ because

$$\mathbb{E}(\mathbb{E}(\varepsilon_{it}|x_{i1},\ldots,x_{iT},\alpha_i)) - \mathbb{E}(\mathbb{E}(\varepsilon_{it-1}|x_{i1},\ldots,x_{iT},\alpha_i)) = 0,$$

so we can base estimation on a set of unconditional moment equations implied by this.

An example is to pool first-differenced observations and do least-squares:

$$\sum_{i=1}^{N}\sum_{t=2}^{T}\Delta x_{it}(\Delta y_{it} - \Delta x_{it}'\beta) = 0,$$

yielding

$$\hat{\beta}_{\mathrm{FD}} := \left(\sum_{i=1}^{N}\sum_{t=2}^{T}\Delta x_{it}\Delta x_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T}\Delta x_{it}\Delta y_{it}\right).$$

Another example

Treatment evaluation with self selection into treatment.

Potential-outcome notation:

$$y_i = y_i^1 \, x_i + y_i^0 \, (1 - x_i) = y_i^0 + (y_i^1 - y_i^0) \, x_i.$$

We wish to learn

$$\begin{aligned}
\mathbb{E}(y_i^1 - y_i^0 | x_i = 1) =& \mathbb{E}(y_i^1 | x_i = 1) - \mathbb{E}(y_i^0 | x_i = 1) \\
=& \mathbb{E}(y_i^1 | x_i = 1) - \mathbb{E}(y_i^0 | x_i = 0) + \mathbb{E}(y_i^0 | x_i = 0) - \mathbb{E}(y_i^0 | x_i = 1) \\
=& \mathbb{E}(y_i \, | x_i = 1) - \mathbb{E}(y_i \, | x_i = 0) + \mathbb{E}(y_i^0 | x_i = 0) - \mathbb{E}(y_i^0 | x_i = 1)
\end{aligned}$$

the average treatment effect on the treated.

The usual assumption (in a cross section) is that

$$\mathbb{E}(y_i^0 | x_i = 0) = \mathbb{E}(y_i^0 | x_i = 1)$$

as to be able to replace the counterfactual outcome of treated by the observed outcome of non-treated.

Now suppose we have access to a pre-treatment outcome, y_{i0}.

This constitutes a two-wave panel, with

$$y_{i0} = y_{i0}^0, \qquad y_i = y_i^0 + (y_i^1 - y_i^0)\, x_i.$$

(no-one is treated at baseline here.)

Then

$$y_i - y_{i0} = (y_i^0 - y_{i0}^0) + (y_i^1 - y_i^0)\, x_i.$$

Therefore,

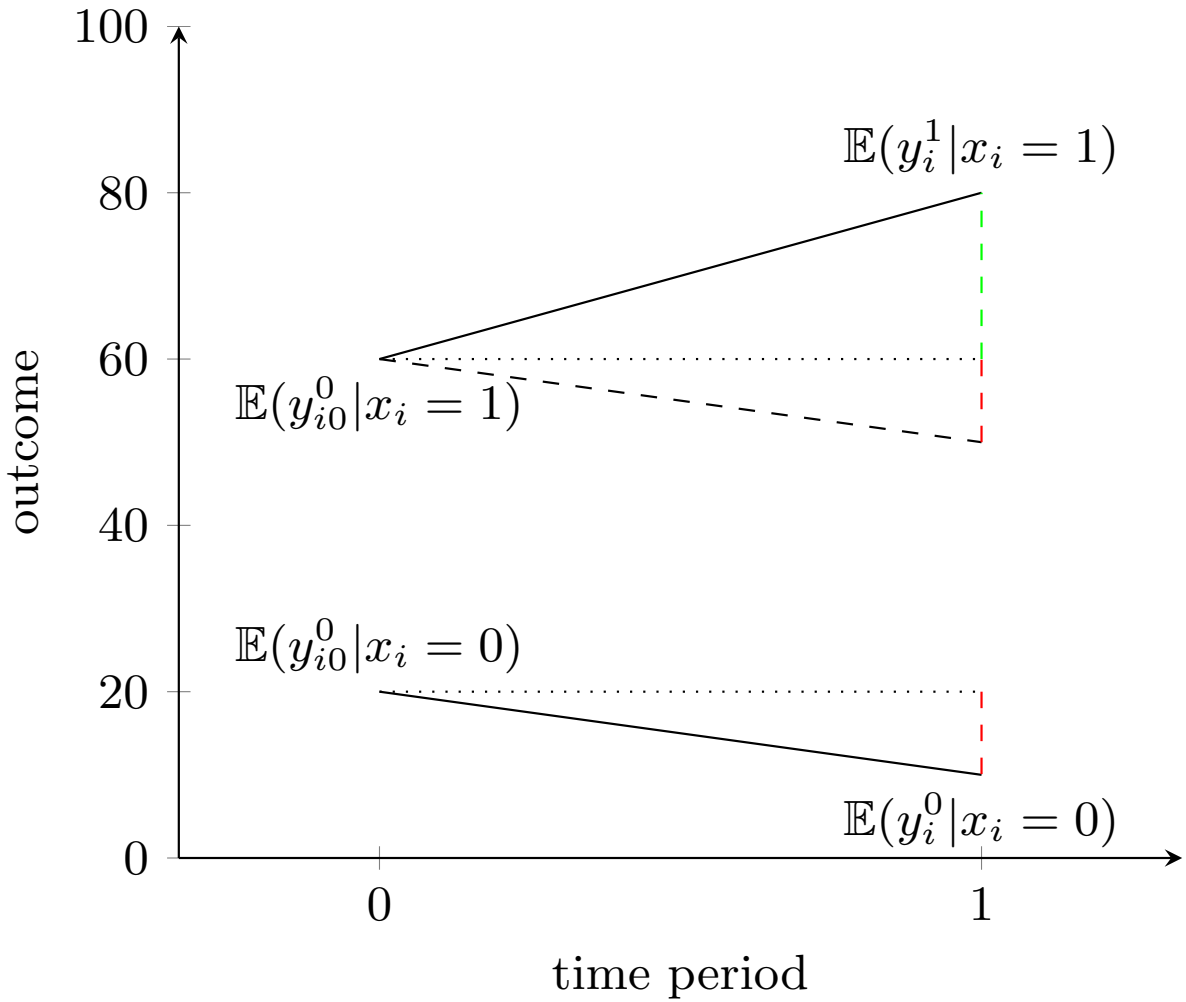
$$\begin{aligned}
\mathbb{E}(y_i - y_{i0}|x_i = 1) =& \mathbb{E}(y_i^0 - y_{i0}^0|x_i = 1) + \mathbb{E}(y_i^1 - y_i^0|x_i = 1), \\
\mathbb{E}(y_i - y_{i0}|x_i = 0) =& \mathbb{E}(y_i^0 - y_{i0}^0|x_i = 0),
\end{aligned}$$

and so

$$\begin{aligned}
\mathbb{E}(y_i^1 - y_i^0|x_i = 1) =& \mathbb{E}(y_i - y_{i0}|x_i = 1) - \mathbb{E}(y_i - y_{i0}|x_i = 0) \\
&+ \mathbb{E}(y_i^0 - y_{i0}^0|x_i = 0) - \mathbb{E}(y_i^0 - y_{i0}^0|x_i = 1)
\end{aligned}$$

and we only require that

$$\mathbb{E}(y_i^0 - y_{i0}^0|x_i = 0) = \mathbb{E}(y_i^0 - y_{i0}^0|x_i = 1).$$


100
80
60
40
20
0
outcome
$\mathbb{E}(y_i^1|x_i = 1)$
$\mathbb{E}(y_{i0}^0|x_i = 1)$
$\mathbb{E}(y_{i0}^0|x_i = 0)$
$\mathbb{E}(y_i^0|x_i = 0)$
0
1
time period

The identifying assumption here is a parallel-trend assumption.

The above is equivalent to a specification of the form

$$\begin{aligned} y_{i0} &= \alpha_i + \delta_0 + \varepsilon_{i0} \\ y_i &= \alpha_i + \delta_1 + x_i\beta + \varepsilon_i \end{aligned}$$

with $\mathbb{E}(\varepsilon_{i0}|x_i, \alpha_i) = \mathbb{E}(\varepsilon_i|x_i, \alpha_i) = 0$.

Then

$$(y_i - y_{i0}) = (\delta_1 - \delta_0) + x_i\beta + (\varepsilon_i - \varepsilon_{i0}),$$

which we estimate by fitting a standard linear model to first-differenced data:

$$\frac{\mathbb{E}(x_i\,\Delta y_i) - \mathbb{E}(x_i)\,\mathbb{E}(\Delta y_i)}{\mathbb{E}(x_i^2) - \mathbb{E}(x_i)^2} = \mathbb{E}(\Delta y_i|x_i = 1) - \mathbb{E}(\Delta y_i|x_i = 0) = \beta.$$

Usually called difference in differences.

Outline

Within-group estimation

Generic formulation:

$$y_{it} = x_{it}'\beta + \alpha_i + \varepsilon_{it}, \qquad \mathbb{E}(\varepsilon_{it}|x_{i1}, \ldots, x_{iT}, \alpha_i) = 0.$$

The requirement on the error is a strict-exogeneity condition.

Note that we do not restrict $\mathbb{E}(\alpha_i|x_{i1}, \ldots, x_{iT})$. Dependence of arbitrary form is allowed.

Can vectorize to get

$$y_i = X_i\beta + \iota_T\alpha_i + \varepsilon_i$$

for

$$y_i := \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}, \qquad X_i := \begin{pmatrix} x_{i1}' \\ \vdots \\ x_{iT}' \end{pmatrix}, \qquad \varepsilon_i := \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix}.$$

Let

$$D := \begin{pmatrix} -1 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \cdots & \vdots \\ 0 & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix}$$

by the $(T-1) \times T$ matrix that transforms data in levels into first differences.

Take first-differences to sweep out the fixed effect

$$Dy_i = DX_i\beta + D\varepsilon_i.$$

Note that, if $\varepsilon_i | X_i \sim (0, \sigma^2 I_T)$, then

$$D\varepsilon_i | X_i \sim (0, \sigma^2 DD');$$

first-differenced errors have a moving-average structure.

The generalized least-squares estimator of β solves

$$\sum_{i=1}^{N} X_i' D' (DD')^{-1} D(y_i - X_i \beta) = 0.$$

Note that

$$M := D'(DD')^{-1} D = I_T - \frac{\iota_T \iota_T'}{T}$$

so that

$$My_i = y_i - \iota_T \frac{\iota_T' y_i}{T} = y_i - \iota_T \bar{y}_i.$$

M transforms data in levels into deviations from within-group means.

The corresponding within-group estimator is

$$\hat{\beta}_{\text{WG}} := \left(\sum_{i=1}^{N} X_i' M X_i \right)^{-1} \left(\sum_{i=1}^{N} X_i' M y_i \right).$$

The need for within-group variation

Note that we require that

$$\sum_{i=1}^{N} X_i' M X_i = \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'$$

has full rank.

Otherwise the estimator is not unique.

This is the usual no-colinearity condition

Here this means that regressors need to have variation within (at least some) groups

Do not include a constant term or things such as race, gender, etc. that do not vary within units.

WG via a dummy-variable regression

We can stack the equations

$$y_i = X_i\beta + \iota_T\alpha_i + \varepsilon_i$$

over units to get

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} \beta + \begin{pmatrix} \iota_T & 0 & \dots & 0 \\ 0 & \iota_T & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \iota_T \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

which can be written compactly as as

$$y = X\beta + B\alpha + \varepsilon$$

for $B = I_N \otimes \iota_T$.

We can then treat α as a (large) parameter and estimate it jointly with β through least squares.

Standard partitioned-regression results applied to

$$y = X\beta + B\alpha + \varepsilon$$

imply that

$$\hat{\beta}_{\text{FE}} = (X'(I_{NT} - B(B'B)^{-1}B')X)^{-1}(X'(I_{NT} - B(B'B)^{-1}B')y) = \hat{\beta}_{\text{WG}}$$

because

$$I_{NT} - B(B'B)^{-1}B' = I_N \otimes M.$$

Indeed,

$$B'B = (I_N \otimes \iota_T)'(I_N \otimes \iota_T) = (I_N \otimes \iota_T')(I_N \otimes \iota_T) = I_N \otimes \iota_T'\iota_T = I_N \otimes T$$
$$BB' = (I_N \otimes \iota_T)(I_N \otimes \iota_T)' = (I_N \otimes \iota_T)(I_N \otimes \iota_T') = I_N \otimes \iota_T\,\iota_T'$$

and so

$$I_{NT} - B(B'B)^{-1}B' = (I_N \otimes I_T) - (I_N \otimes \iota_T\iota_T'/T) = (I_N \otimes M).$$

Usually refered to as fixed-effect or least-squares dummy-variable estimator.

WG via a control-function regression

Consider pooling the data and regressing y_{it} on a constant, x_{it} and $\bar{x}_i$.

By partitioned regression this amounts to regressing y_{it} on the residual from a regression of x_{it} on a constant and $\bar{x}_i$:

$$\min_{\delta_1,\delta_2} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \delta_1 - \delta_2 \bar{x}_i)^2.$$

Applying the elementary least-squares formulae here yields

$$\hat{\delta}_2 = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} (\bar{x}_i - \bar{x}) x_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T} (\bar{x}_i - \bar{x}) \bar{x}_i} = \frac{\sum_{i=1}^{N} (\bar{x}_i - \bar{x}) \bar{x}_i}{\sum_{i=1}^{N} (\bar{x}_i - \bar{x}) \bar{x}_i} = 1$$

and

$$\hat{\delta}_1 = \bar{x} - \hat{\delta}_2 \bar{x} = 0.$$

So the relevant residual is simply $x_{it} - \bar{x}_i$, which again yields within-groups.

Statistical properties

If

$$\varepsilon_i \sim (0, \sigma^2 I_T)$$

then within-groups is the optimal generalized least-squares estimator.

Consequently, it is best linear unbiased by the Gauss-Markov theorem.

It remains unbiased for more general error structures

$$\varepsilon_i \sim (0, \Sigma)$$

although it will no longer be optimal in the above sense.

WG is nonetheless the workhorse estimator for the linear panel model (at least, provided that strict exogeneity holds!).

Be sure to use suitably robust standard errors for inference.

Consistency

The usual asymptotic framework has $N \to \infty$ and T held fixed, and we presume random sampling in the cross section.

In this case WG is a least-squares estimator on system of T equations.

We have

$$\hat{\beta}_{\text{WG}} - \beta = \left(\sum_{i=1}^{N} X_i' M X_i \right)^{-1} \left(\sum_{i=1}^{N} X_i' M \varepsilon_i \right) \xrightarrow{p} 0$$

provided that

- The matrix $A := \mathbb{E}(X_i' M X_i)$ exists and has maximal rank, and
- $\mathbb{E}(\varepsilon_i | X_i, \alpha_i) = 0$ holds.

Limit distribution

Further, assuming that

- the matrix $C := \mathbb{E}(X_i' M \Sigma M X_i)$ exists,

we have

$$\sqrt{N}(\hat{\beta}_{\mathrm{WG}} - \beta) \xrightarrow{d} \mathbf{N}(0, A^{-1} C A^{-1}).$$

When $\varepsilon \sim (0, \sigma^2 I_T)$ we have

$$C = \sigma^2 A$$

and

$$\sqrt{N}(\hat{\beta}_{\mathrm{WG}} - \beta) \xrightarrow{d} \mathbf{N}(0, \sigma^2 A^{-1})$$

but we will usually not presume this.

Variance estimation

For inference we need consistent estimators of A and C.

For A we use

$$\frac{1}{N}\sum_{i=1}^{N} X_i' M X_i.$$

For C, first consider the simple case where $C = \sigma^2 A$.

Write

$$\hat{\varepsilon}_i = M(y_i - X_i \hat{\beta}_{\mathrm{WG}}).$$

The naive plug-in estimator would use

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} \frac{\hat{\varepsilon}_i' \hat{\varepsilon}_i}{T} = \frac{1}{N}\sum_{i=1}^{N} \frac{\varepsilon_i' M \varepsilon_i}{T} + o_p(1).$$

With $\varepsilon_i \sim (0, \sigma^2 I_T)$ we have

$$\mathbb{E}(\varepsilon_i' M \varepsilon_i) = \sigma^2 \mathrm{trace}(M) = \sigma^2 (T-1).$$

Hence,

$$\hat{\sigma}^2 \xrightarrow{p} \sigma^2 \frac{T-1}{T} = \sigma^2 - \frac{\sigma^2}{T}.$$

The inconsistency of $\hat{\sigma}^2$ is a consequence of the estimation noise in the fixed effects.

A degrees-of-freedom correction solves the problem:

$$\tilde{\sigma}^2 := \frac{T}{T-1}\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} \frac{\varepsilon_i' M \varepsilon_i}{T-1}.$$

A robust estimator of C is

$$\frac{1}{N}\sum_{i=1}^{N} X_i' M \hat{\varepsilon}_i \hat{\varepsilon}_i' M X_i.$$

This is often called cluster-robust variance estimator

It handles general forms of both heteroskedasticity and serial correlation.

Traditional heteroskedasticity robust variance estimator

The view of WG as a dummy-variable least-squares regression may suggest using a traditional (cross-sectional) 'White'-type variance formula to deal with heteroskedasticity.

This amounts to estimating C by

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'\hat{\varepsilon}_{it}^2.$$

This estimator is inconsistent.

Indeed,

$$\hat{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i + o_p(1)$$

and so

$$\mathbb{E}(\hat{\varepsilon}_{it}^2|X_i) = \mathbb{E}((\varepsilon_{it}-\bar{\varepsilon}_i)^2|X_i) = \mathbb{E}(\varepsilon_{it}^2|X_i) + O(T^{-1}).$$

Comment: Specification testing

The above also highlights that usual specification tests developed for cross-sectional or time series problems, such as tests for heteroskedasticity or serial correlation should not be used here.

If desired, panel data alternatives for some of these tests have been developed and should be used instead.

Outline

The random-effect model

The classical random-effect approach for

$$y_i = X_i\beta + \iota_T\alpha_i + \varepsilon_i$$

assumes that

$$\varepsilon_i|X_i, \alpha_i \sim (0, \sigma^2 I_T), \qquad \alpha_i|X_i \sim (0, \gamma^2 I_N);$$

here a constant term (and if need be other variables that do not vary over time) is (and can be) included as regressor.

One implication is that

$$\mathbb{E}(y_i|X_i) = X_i\beta$$

so that a pooled least-squares regression is consistent.

This is not suitable when we view α_i as an unobserved confounding factor!

Estimation

Here, the error satisfies

$$u_i = \iota_T \alpha_i + \varepsilon_i \sim (0, \Omega)$$

for

$$\Omega = \begin{pmatrix} \sigma^2 + \gamma^2 & \gamma^2 & \dots & \gamma^2 \\ \gamma^2 & \sigma^2 + \gamma^2 & \dots & \gamma^2 \\ \vdots & \ddots & \ddots & \vdots \\ \gamma^2 & \gamma^2 & \dots & \sigma^2 + \gamma^2 \end{pmatrix} = \sigma^2 I_T + \gamma^2 \iota_T \iota_T'.$$

The (infeasible) generalizes least-squares estimator thus is

$$\left(\sum_{i=1}^{N} X_i' \Omega^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{N} X_i' \Omega^{-1} y_i \right).$$

To construct a feasible version of the GLS estimator we can use residuals from a pooled least-squares regression, say $\hat{u}_{it}$.

We estimate the diagonal entries of Ω as

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{u}_{it}^2 \xrightarrow{p} \sigma^2 + \gamma^2,$$

and the off-diagonal entries of Ω as

$$\frac{1}{NT(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s\neq t}\hat{u}_{it}\hat{u}_{is} \xrightarrow{p} \gamma^2.$$

The random-effect estimator then is

$$\hat{\beta}_{\text{RE}} := \left(\sum_{i=1}^{N} X_i'\hat{\Omega}^{-1}X\right)^{-1}\left(\sum_{i=1}^{N} X_i'\hat{\Omega}^{-1}y\right).$$

Correlated random effects

We can model the dependence between α_i and the x_{it}.

A simple example would be

$$\alpha_i | X_i \sim (\bar{x}_i' \delta, \gamma^2).$$

This corresponds to

$$\alpha_i = \bar{x}_i' \delta + \eta_i, \qquad \eta_i \sim (0, \gamma^2).$$

Substitution into the model yields

$$y_{it} = x_{it}' \beta + \alpha_i + \varepsilon_{it} = x_{it}' \beta + \bar{x}_i' \delta + (\varepsilon_{it} + \eta_i).$$

Can estimate this by random effects, as before.

Note that pooled estimation yields the fixed-effect estimator!

Outline

The strict-exogeneity assumption

$$\mathbb{E}(\varepsilon_{it}|x_{i1},\ldots,x_{iT},\alpha_i)=0$$

has been at the heart of our developments so far.

Often too strong to maintain.

It rules out dynamics and, more generally, feedback from current outcomes to future regressors.

These would lead to bias and inconsistency in the within-group estimator.

A dynamic model

Consider the basic case where

$$y_{it} = \alpha_i + \rho y_{it-1} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \text{i.i.d.}(0, \sigma^2),$$

and we observe an initial observation y_{i0}.

The within-group estimator is based on the presumption that

$$\mathbb{E}((y_{it-1} - \bar{y}_{i-})\varepsilon_{it}) = 0.$$

However,

$$\begin{aligned}
\mathbb{E}((y_{it-1} - \bar{y}_{i-})\varepsilon_{it}) &= -\mathbb{E}(\bar{y}_{i-}\varepsilon_{it}) \\
&= -\frac{1}{T}\sum_{s=0}^{T-1}\mathbb{E}(y_{is}\varepsilon_{it}) = -\frac{1}{T}\sum_{s=t}^{T-1}\mathbb{E}(y_{is}\varepsilon_{it}) \\
&= -\frac{\mathbb{E}(y_{it}\varepsilon_{it})}{T} - \frac{\mathbb{E}(y_{it+1}\varepsilon_{it})}{T} - \cdots - \frac{\mathbb{E}(y_{iT-1}\varepsilon_{it})}{T} \\
&= -\frac{\sigma^2}{T} - \frac{\rho\sigma^2}{T} - \cdots - \frac{\rho^{T-t}\sigma^2}{T}.
\end{aligned}$$

The de-meaning introduces endogeneity bias of its own kind.

In the current example, we have

$$\mathbb{E}(\varepsilon_{it}|y_{i0},\ldots,y_{it-1},\alpha_i)=0.$$

(Note that this implies that errors are serially uncorrelated.)

Recursive substitution gives

$$y_{it}=\alpha_i+\rho y_{it-1}+\varepsilon_{it}=\frac{\alpha_i}{1-\rho}(1-\rho^t)+\rho^t y_{i0}+(\varepsilon_{it}+\rho\varepsilon_{it-1}+\cdots+\rho^{t-1}\varepsilon_{i1}).$$

Taking first-differences gives

$$\Delta y_{it}=\rho\Delta y_{it-1}+\Delta\varepsilon_{it}.$$

Pooled least-squares on first differences would again be inconsistent because

$$\mathbb{E}(\Delta y_{it-1}\Delta\varepsilon_{it})=-\mathbb{E}(y_{it-1}\Delta\varepsilon_{it})=-\mathbb{E}(y_{it-1}\varepsilon_{it-1})=-\sigma^2\neq 0.$$

However, because errors are uncorrelated over time,

$$\mathbb{E}(y_{it-2}\Delta\varepsilon_{it-1})=\mathbb{E}(y_{it-3}\Delta\varepsilon_{it-1})=\cdots=\mathbb{E}(y_{i0}\Delta\varepsilon_{it-1})=0$$

while, for $j=2,3,\ldots$

$$\mathbb{E}(y_{it-j}\Delta y_{it-1})\neq 0;$$

these terms do depend on the distribution of (α_i,y_{i0}) (!).

GMM estimator

Note that we do not make assumptions on the distribution of $y_{i0}|\alpha_i$; initial conditions can have heterogeneous distributions (across agents) that need not equal the steady-state distribution of the markov process.

This is useful as stationarity can be a restrictive assumption in short panels.

The above argument gives rise to a set of sequential moment conditions:

$$\mathbb{E}\left(\begin{pmatrix} y_{it-2} \\ y_{it-3} \\ \vdots \\ y_{i0} \end{pmatrix}(\Delta y_{it} - \rho \Delta y_{it-1})\right) = 0$$

for each $t = 2, \ldots, T$.

This gives a total of $T(T-1)/2$ moment equations that can be combined through a conventional GMM procedure.

They do not rely on homoskedasticity.

Let

$$Z_i := \begin{pmatrix} y_{i0} & 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & y_{i0} & y_{i1} & \dots & 0 & \dots & 0 \\ \vdots & & & \ddots & & & \vdots \\ 0 & 0 & 0 & \dots & y_{i0} & \dots & y_{i(T-2)} \end{pmatrix}.$$

Then we use the empirical moments

$$\frac{1}{N}\sum_{i=1}^{N} Z_i'(\Delta y_i - \rho \Delta y_-) = 0.$$

Stacking blocks across individuals to get

$$Z'\Delta y_- := (Z_1', \dots, Z_N') \begin{pmatrix} \Delta y_{1-} \\ \vdots \\ \Delta y_{N-} \end{pmatrix}, \quad Z'\Delta y := (Z_1', \dots, Z_N') \begin{pmatrix} \Delta y_1 \\ \vdots \\ \Delta y_N \end{pmatrix},$$

the estimator for a given weight matrix A is

$$\hat{\rho}_{\text{IV}} = \left(\Delta y_-' Z A Z' \Delta y_-\right)^{-1} \left(\Delta y_-' Z A Z' \Delta y\right).$$

This GMM estimator is consistent and asymptotically-normal under fixed-T asymptotics (under usual regularity conditions).

In

$$\hat{\rho}_{\text{IV}} = \left(\Delta y'_{-} Z A Z' \Delta y_{-}\right)^{-1} \left(\Delta y'_{-} Z A Z' \Delta y\right)$$

the optimal choice for the weight matrix is

$$A_{\text{opt}} = \left(\frac{1}{N} \sum_{i=1}^{N} Z'_i \Delta\hat{\varepsilon}_i \Delta\hat{\varepsilon}'_i Z_i \right)^{-1}$$

where $\Delta\hat{\varepsilon}_i$ are one-step GMM residuals.

The asymptotic variance of the two-step estimator can be estimated by

$$\left(\Delta y'_{-} Z A_{\text{opt}} Z' \Delta y_{-}\right)^{-1}.$$

Comment: Long panels

The GMM estimator has non-negligible bias in long panels.

(Under homoskedasticity) we have, as $N, T \to \infty$ with $N/T \to c$ for some finite constant $c \in (0, +\infty)$,

$$\sqrt{NT}\left(\hat{\rho}_{\text{IV}} - \rho + \frac{1+\rho}{N}\right) \xrightarrow{d} \mathbf{N}(0, 1-\rho^2).$$

Compare this to WG under the same conditions:

$$\sqrt{NT}\left(\hat{\rho}_{\text{WG}} - \rho + \frac{1+\rho}{T}\right) \xrightarrow{d} \mathbf{N}(0, 1-\rho^2).$$

Here it would suffice to use

$$\hat{\rho}_{\text{WG}} + \frac{1+\hat{\rho}_{\text{WG}}}{T},$$

a bias-corrected within-group estimator.

Accommodating serial dependence

The approach can be modified to

$$\mathbb{E}(\varepsilon_{it}|y_{i0},\ldots,y_{it-p},\alpha_i)=0.$$

for some $1 \leq p \leq t$.

This allows for dependence between ε_{it} and ε_{it-j} for $1 \leq j < p$.

A simple example would be a moving-average process, e.g.,

$$\varepsilon_{it}=\eta_{it}+\theta\eta_{it-1}, \qquad \eta_{it} \sim \text{i.i.d.}(0,\sigma_\eta^2).$$

The moment restriction is not compatible with an autoregressive process, however, such as

$$\varepsilon_{it}=\varrho\varepsilon_{it-1}+\eta_{it}, \qquad \eta_{it} \sim \text{i.i.d.}(0,\sigma_\eta^2).$$

as correlation decays exponentially in t.

Higher-order dynamics

The above discussion is straightforward to adapt to

$$y_{it} = \alpha_i + \rho_1 y_{it-1} + \rho_2 y_{it-2} + \cdots + \rho_p y_{it-p} + \varepsilon_{it}$$

for some p provided that the serial dependence in the time-varying errors is suitably restricted.

Identifying power of covariates

An interesting generalization is

$$y_{it} = \alpha_i + \rho y_{it-1} + x_{it}'\beta + \varepsilon_{it}$$

under the assumption that

$$\mathbb{E}(\varepsilon_{it}|x_{i1}, \ldots, x_{iT}, \alpha_i) = 0.$$

Here,

- regressors x_{it} are strictly exogenous,
- errors ε_{it} may be serially correlated,
- lagged values y_{it-1} are effectively (treated as) endogenous.

Exogeneity yields moment conditions of the form

$$\mathbb{E}\left(\begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iT} \end{pmatrix} (\Delta y_{it} - \rho \Delta y_{it-1} - \Delta x_{it}'\beta)\right) = 0.$$

General feedback problems

Feedback of arbitrary form such as in

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

with

$$\mathbb{E}(\varepsilon_{it}|x_{i1}, \ldots, x_{it}, \alpha_i) = 0.$$

can be handled with GMM based on

$$\mathbb{E}\left(\begin{pmatrix} x_{it-1} \\ x_{it-2} \\ \vdots \\ x_{i1} \end{pmatrix} (\Delta y_{it} - \Delta x_{it}\beta)\right) = 0.$$

Outline

Additive effects

Models that are nonlinear in common parameters but additive in fixed effects pose no substantial problems.

Take

$$y_{it} = \varphi(x_{it}, \beta) + \alpha_i + \varepsilon_{it}, \qquad \mathbb{E}(\varepsilon_{it} | x_{i1}, \ldots, x_{iT}, \alpha_i) = 0$$

and φ some known function.

Can still remove fixed effects by differencing.

Can again construct a GMM estimator based on, say,

$$\mathbb{E}\left(\begin{pmatrix} x_{iT} \\ \vdots \\ \\ \vdots \\ x_{i1} \end{pmatrix} ((y_{it} - y_{it-1}) - (\varphi(x_{it}, \beta) - \varphi(x_{it-1}, \beta))) \right) = 0.$$

Limited dependent variables

The additive specification will not be suitable in cases where the outcome is limited:

- discrete choice
- count outcomes
- etc.

It is not possible, in general, to separate the problem of inference on β from estimation of α_i.

An approach that estimates α_i jointly with β will, in general, be inconsistent for β when T is treated as fixed.

This is known as the incidental-parameter problem.

Incidental-parameter problem

Take a likelihood framework. Maximum likelihood solves the concentrated problem

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{t=1}^{T} \log f(z_{it}; \theta, \hat{\alpha}_i(\theta)), \quad \hat{\alpha}_i(\theta) = \underset{\alpha_i}{\operatorname{argmax}} \sum_{t=1}^{T} \log f(z_{it}; \theta, \alpha_i).$$

We have

$$\theta_0 = \underset{\theta}{\operatorname{argmax}} \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left(\log f(z_{it}; \theta, \alpha_i(\theta)) \right),$$

with $\alpha_i(\theta) = \underset{\alpha_i}{\operatorname{argmax}} \mathbb{E}(\log f(z_{it}; \theta, \alpha_i))$.

However, with T fixed,

$$\hat{\theta} \xrightarrow{p} \underset{\theta}{\operatorname{argmax}} \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{n} \mathbb{E} \left(\log f(z_{it}; \theta, \hat{\alpha}_i(\theta)) \right)$$

and $\hat{\alpha}_i(\theta) \overset{p}{\nrightarrow} \alpha_i(\theta)$.

Multiplicative effects

One case where positive results can be obtained is in models of the form

$$y_{it} = \varphi(x_{it}, \beta)\, \alpha_i\, \varepsilon_{it}, \qquad \mathbb{E}(\varepsilon_{it} | x_{i1}, \ldots, x_{iT}, \alpha_i) = 1$$

and φ some known function.

One popular example is a count-data regression:

$$\mathbb{E}(y_{it} | x_{i1}, \ldots, x_{iT}, \alpha_i) = \exp(x_{it}'\beta)\, \alpha_i.$$

Another example would be a binary-choice model with

$$\mathbb{P}(y_{it} = 1 | x_{i1}, \ldots, x_{iT}, \alpha_i) = F(x_{it}'\beta)\, \alpha_i$$

for some cdf F and $\alpha \in (0, 1)$.

This setting is peculiar because

$$\mathbb{E}\left(\frac{y_{it}}{\varphi(x_{it}, \beta)} \middle| x_{i1}, \ldots, x_{iT}, \alpha_i \right) = \alpha_i$$

for all $t = 1, \ldots, T$.

Therefore, we can 'difference-out' the fixed effects by using

$$\mathbb{E}\left(\frac{y_{it}}{\varphi(x_{it}, \beta)} - \frac{y_{it-1}}{\varphi(x_{it-1}, \beta)} \middle| x_{i1}, \ldots, x_{iT}, \alpha_i \right) = 0$$

for all $t = 2, \ldots, T$.

We can also take deviations from within-group means:

$$\mathbb{E}\left(\frac{y_{it}}{\varphi(x_{it}, \beta)} - \frac{\sum_j y_{ij}}{\sum_j \varphi(x_{ij}, \beta)} \middle| x_{i1}, \ldots, x_{iT}, \alpha_i \right) = 0$$

One unconditional moment equation arising from this last argument is

$$\mathbb{E}\left(x_{it}\left(y_{it}-\left(\frac{\sum_j y_{ij}}{\sum_j \varphi(x_{ij},\beta)}\right)\varphi(x_{it},\beta)\right)\right)=0$$

The estimator based on this is often called the pseudo-poisson estimator.

The above moment condition is equal to the (profiled/concentrated) score equation for β if $y_{it}|x_{i1},\ldots,x_{iT},\alpha_i$ is assumed to be poisson distribution with mean $\varphi(x_{it},\beta)\,\alpha_i$, and all parameters are estimated jointly by standard maximum likelihood.

The pseudo-poisson estimator therefore does not need the data to be poisson distributed. It does not require the data to be count data neither; they can be continuous with a mass point at zero. You do need to use robust standard errors for inference.

Do not use pseudo-poisson when regressors are not strictly exogenous! On the other hand, you can still construct a differencing-based estimator based on sequential moment restrictions.

Censored regression

Now take

$$y_{it} = \max\{\eta_i + x_{it}\theta + \varepsilon_{it}, 0\}$$

with stationary errors that are independent of regressors, with CDF F.

Take a two-wave panel for simplicity.

Here, only units for which no censoring occurs are informative about θ.

Conditional on y_{i1} and y_{i2} both being uncensored,

$$\mathbb{E}(\Delta y_i - \Delta x_i\theta | x_i, y_{i1} > 0, y_{i2} > 0)$$

equals

$$\mathbb{E}(\Delta y_i - \Delta x_i\theta | x_i, \varepsilon_{i1} > -\eta_i - x_{i1}\theta, \varepsilon_{i2} > -\eta_i - x_{i2}\theta)$$

and this is non-zero, in general.

This is so because the truncated error distributions are different across time.

The moment condition can be restored by artificially censoring further.

Indeed,

$$\mathbb{E}(\Delta\varepsilon_i | \varepsilon_{i1} > \max\{-\eta_i - x_{i1}\theta, -\eta_i - x_{i2}\theta\}, \varepsilon_{i2} > \max\{-\eta_i - x_{i1}\theta, -\eta_i - x_{i2}\theta\})$$

is zero by stationarity.

The errors are unobserved but

$$\begin{aligned}
\varepsilon_{i1} > -\eta_i - x_{i1}\theta &\Leftrightarrow y_{i1} > 0 \\
\varepsilon_{i1} > -\eta_i - x_{i2}\theta &\Leftrightarrow y_{i1} > -\Delta x_i\theta,
\end{aligned}$$

and, similarly,

$$\begin{aligned}
\varepsilon_{i2} > -\eta_i - x_{i1}\theta &\Leftrightarrow y_{i2} > \Delta x_i\theta \\
\varepsilon_{i2} > -\eta_i - x_{i2}\theta &\Leftrightarrow y_{i2} > 0.
\end{aligned}$$

We thus consider a trimmed least-squares estimator

$$\min_\theta \sum_{i=1}^N (\Delta y_i - \Delta x_i\theta)^2 \{y_{i1} > \max\{0, -\Delta x_i\theta\}\} \{y_{i2} > \max\{0, \Delta x_i\theta\}\}$$

Logistic regression

A simple two-period logit model has

$$\mathbb{P}(y_{i1}=1|\alpha_i)=\frac{1}{1+e^{-\alpha_i}}=F(\alpha_i),\quad \mathbb{P}(y_{i2}=1|\alpha_i)=\frac{1}{1+e^{-(\alpha_i+\beta)}}=F(\alpha_i+\beta).$$

Here, β is the log-odds ratio.

The log-likelihood is

$$\sum_{i=1}^{N} y_{i1}\log F(\alpha_i\quad\quad)+(1-y_{i1})\log(1-F(\alpha_i\quad\quad))$$
$$+\sum_{i=1}^{N} y_{i2}\log F(\alpha_i+\beta)+(1-y_{i2})\log(1-F(\alpha_i+\beta)).$$

To profile-out the fixed effects, note that we have four types of units in the data:

- $y_{i1}=0$, $y_{i2}=1$ (movers in)
- $y_{i1}=1$, $y_{i2}=0$ (movers out)
- $y_{i1}=1$, $y_{i2}=1$ (stayers in)
- $y_{i1}=0$, $y_{i2}=0$ (stayers out)

The score equation for α_i is

$$(y_{i1} + y_{i2}) - (F(\alpha_i) + F(\alpha_i + \beta)) = 0.$$

- If $y_{i1} = 0$, $y_{i2} = 1$ (movers in) this is $1 - F(\alpha_i) - F(\alpha_i + \beta) = 0$.
- If $y_{i1} = 1$, $y_{i2} = 0$ (movers out) this is $1 - F(\alpha_i) - F(\alpha_i + \beta) = 0$.
- If $y_{i1} = 1$, $y_{i2} = 1$ (stayers in) this is $2 - F(\alpha_i) - F(\alpha_i + \beta) = 0$.
- If $y_{i1} = 0$, $y_{i2} = 0$ (stayers out) this is $- F(\alpha_i) - F(\alpha_i + \beta) = 0$.

and so

- for movers
$$\hat{\alpha}_i(\beta) = -\beta/2;$$
- for stayers
$$\hat{\alpha}_i(\beta) = \pm\infty.$$

Stayers do not carry information about β, so do not contribute to the profile log-likelihood.

Let $\Delta y_i = y_{i2} - y_{i1}$.

Movers have $\Delta y_i \in \{-1, 1\}$.

The profile log-likelihood is

$$2\sum_{i=1}^{n}\{\Delta y_i = -1\}\log F(-\beta/2) + \{\Delta y_i = \quad 1\}\log F(\beta/2).$$

The profile-score equation is

$$\sum_{i=1}^{N}\{\Delta y_i = \quad 1\}\left(1 - F(\beta/2)\right) - \{\Delta y_i = -1\}F(\beta/2) = 0.$$

With $n_{01} = \sum_{i=1}^{N}\{\Delta y_i = 1\}$ and $n_{10} = \sum_{i=1}^{N}\{\Delta y_i = -1\}$ the score root is

$$\hat{\beta} = 2F^{-1}\left(\frac{n_{01}}{n_{10} + n_{01}}\right) = 2F^{-1}\left(\frac{1}{1 + {}^{n_{10}}/_{n_{01}}}\right) \xrightarrow{p} 2F^{-1}\left(\frac{1}{1 + e^{-\beta}}\right) = 2\beta$$

and so maximum-likelihood is inconsistent.

Here we have used that

$$\text{plim}_{N\to\infty}\frac{n_{10}}{n_{01}} = \frac{\mathbb{E}(\mathbb{P}(\Delta y_i = -1|\alpha_i))}{\mathbb{E}(\mathbb{P}(\Delta y_i = \quad 1|\alpha_i))} = e^{-\beta},$$

which follows from the observation that

$$\mathbb{E}(\mathbb{P}(\Delta y_i = -1|\alpha_i)) = \mathbb{E}\left(\frac{1}{1+e^{-\alpha_i}}\frac{e^{-(\alpha_i+\beta)}}{1+e^{-(\alpha_i+\beta)}}\right),$$

and

$$\mathbb{E}\left(\mathbb{P}(\Delta y_i = \quad 1|\alpha_i)\right) = \mathbb{E}\left(\frac{e^{-\alpha_i}}{1+e^{-\alpha_i}}\frac{1}{1+e^{-(\alpha_i+\beta)}}\right),$$

so that

$$\mathbb{E}\left(\mathbb{P}(\Delta y_i = -1|\alpha_i)\right) = e^{-\beta}\ \left(\mathbb{E}\left(\mathbb{P}(\Delta y_i = \quad 1|\alpha_i)\right)\right).$$

Conditional likelihood

Consider again

$$\mathbb{P}(y_{it} = 1 | x_{i1}, \ldots, x_{iT}, \alpha_i) = \frac{1}{1 + e^{-(\alpha_i + x_{it}\beta)}} = F(\alpha_i + x_{it}\beta)$$

and maintain a two-wave panel.

A sufficient statistic here is (any monotone function of) the sum $y_{i1} + y_{i2}$.

Recall that the likelihood contribution of stayers does not contain information on β.

Relevant case is, therefore, $y_{i1} + y_{i2} = 1$. These are movers in and out of the waves.

First,

$$\mathbb{P}(y_{i1} = 0, y_{i2} = 1 | x_i, y_{i1} + y_{i2} = 1, \alpha_i)$$

is equal to

$$\frac{1}{1 + \frac{\mathbb{P}(y_{i1}=1, y_{i2}=0|x_i,\alpha_i)}{\mathbb{P}(y_{i1}=0, y_{i2}=1|x_i,\alpha_i)}}.$$

Now,

$$\mathbb{P}(y_{i1} = 0, y_{i2} = 1 | x_i, \alpha_i) = \frac{e^{-(\alpha_i + x_{i1}\beta)}}{1 + e^{-(\alpha_i + x_{i1}\beta)}} \frac{}{1 + e^{-(\alpha_i + x_{i2}\beta)}}$$

$$\mathbb{P}(y_{i1} = 1, y_{i2} = 0 | x_i, \alpha_i) = \frac{1}{1 + e^{-(\alpha_i + x_{i1}\beta)}} \frac{e^{-(\alpha_i + x_{i2}\beta)}}{1 + e^{-(\alpha_i + x_{i2}\beta)}}$$

and so

$$\frac{\mathbb{P}(y_{i1} = 1, y_{i2} = 0 | x_i)}{\mathbb{P}(y_{i1} = 0, y_{i2} = 1 | x_i, \alpha_i)} = \frac{e^{-(\alpha_i + x_{i2}\beta)}}{e^{-(\alpha_i + x_{i1}\beta)}} = e^{-(x_{i2} - x_{i1})\beta}.$$

Therefore,

$$\mathbb{P}(y_{i1} = 0, y_{i2} = 1 | x_i, y_{i1} + y_{i2} = 1, \alpha_i) = \frac{1}{1 + e^{-(x_{i2} - x_{i1})\beta}} = F((x_{i2} - x_{i1})\beta)$$

Note that this implies that Δy_i conditional on $\Delta y_i \neq 0$ is Bernoulli with succes probability

$$\mathbb{P}(\Delta y_i = 1 | x_i, \Delta y_i \neq 0) = F(\Delta x_i \beta).$$

This is a logistic regression (for the subpanel of movers) in first differences.

The conditional log-likelihood is

$$\sum_{i=1}^{N} \{\Delta y_i = \quad 1\} \log(F(\Delta x_i \beta)) + \{\Delta y_i = -1\} \log(1 - F(\Delta x_i \beta)).$$

The score is

$$\sum_{i=1}^{N} \Delta x_i (\{\Delta y_i = 1\} - F(\Delta x_i \beta)) = 0$$

and is clearly unbiased.

The assumption of F being logistic is important here. Other choices do not lead to sufficiency.

Outline

Rectangular-array asymptotics

Potential for identification in short panels is limited.

Existence of moment conditions is very specific to the specification and the parameter of interest.

Also, asymptotics that treat the length of the panel as fixed do not suit all problems.

In increasingly many empirical settings the length of the panel is statistically informative about individual-specific parameters.

Asymptotics where N and T grow large at the same rate—i.e., such that N/T converges to a finite constant—give an accurate reflection of the sampling behavior here.

Here, the incidental-parameter problem manifests itself as an asymptotic-bias problem.

An example: Linear model

Stripped-down version of linear model is

$$y_{it} \sim \mathbf{N}(\alpha_i, \theta).$$

Here,

$$\hat{\theta} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \bar{y}_i)^2$$

has bias and variance

$$\mathbb{E}(\hat{\theta}) = \theta \left(1 - \frac{1}{T}\right) \qquad \text{var}(\hat{\theta}) = \frac{2\theta^2}{NT} \left(1 - \frac{1}{T}\right).$$

As $N, T \to \infty$ with $N/T \to c^2$ bias and standard deviation are of the same order and, hence,

$$\sqrt{NT}(\hat{\theta} - \theta) \to \mathbf{N}(-c\,\theta, 2\theta^2),$$

is not centered at zero.

The estimator is consistent but asymptotically biased.

Bias correction

For general nonlinear model, under regularity conditions,

$$\text{plim}_{N\to\infty}\hat{\theta} = \theta + \frac{B}{T} + o(T^{-1}).$$

The leading bias term, B, can be estimated to construct a bias-corrected estimator

$$\hat{\theta} - \frac{\hat{B}}{T}.$$

The bias can be estimated based on analytical formulae using the fixed-effect estimator.

A simple alternative is to use a jackknife.

As an example of a jackknife suppose that individual time series are ergodic and stationary.

Split the data into two (non-overlapping) subpanels of adjacent observations.

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ denote the corresponding estimators, based on $N \times T_1$ and $N \times T_2$ observations, respectively, where $T_1 + T_2 = T$.

Then,

$$\text{plim}_{N\to\infty}\hat{\theta}_1 = \theta + \frac{B}{T_1} + o(T_1^{-1}), \qquad \text{plim}_{N\to\infty}\hat{\theta}_2 = \theta + \frac{B}{T_2} + o(T_2^{-1}).$$

Hence,

$$\bar{\theta} = \frac{T_1\,\hat{\theta}_1 + T_2\,\hat{\theta}_2}{T} \xrightarrow{p} \theta + 2\frac{B}{T} + o(T^{-1}).$$

It follows that the jackknife bias-correction estimator

$$2\hat{\theta} - \bar{\theta}$$

is asymptotically unbiased.

Further, because $\hat{\theta}_1$ and $\hat{\theta}_2$ are asymptotically independent this estimator has the same large-sample variance as $\hat{\theta}$.

As $N, T \to \infty$ with $N/T \to c^2$,

Maximum-likelihood satisfies

$$\sqrt{NT}(\hat{\theta} - \theta) \xrightarrow{d} \mathbf{N}(cB, I_\theta^{-1}).$$

(for I_θ the Fisher information, as usual).

Bias-corrected estimator, $\check{\theta}$, satisfies

$$\sqrt{NT}(\check{\theta} - \theta) \xrightarrow{d} \mathbf{N}(\ 0\ , I_\theta^{-1}).$$

Bias correction justifies the use of conventional inference procedures (test statistics and confidence sets) in rectangular panels.

Results carry over the average marginal effects in nonlinear models.

Numerical implementation

Fixed-effect models have a large number of parameters.

Newton-Raphson type optimization requires the inverse Hessian matrix for all parameters.

Often judged to be computationally difficult or even infeasible.

Not generally true.

The Hessian is sparse. Can be computed efficiently using only inverses of low-dimensional matrices.

The score and Hessian are

$$\begin{pmatrix} \ell_\theta \\ \ell_{\alpha_1} \\ \ell_{\alpha_2} \\ \vdots \\ \ell_{\alpha_N} \end{pmatrix}, \qquad \begin{pmatrix} \ell_{\theta\theta} & \ell_{\theta\alpha_1} & \ell_{\theta\alpha_2} & \cdots & \ell_{\theta\alpha_N} \\ \ell_{\alpha_1\theta} & \ell_{\alpha_1\alpha_1} & 0 & \cdots & 0 \\ \ell_{\alpha_2\theta} & 0 & \ell_{\alpha_2\alpha_2} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \ell_{\alpha_N\theta} & 0 & 0 & \cdots & \ell_{\alpha_N\alpha_N} \end{pmatrix},$$

where the individual components are

$$\ell_\theta = \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{\partial \log f(z_{it}|\theta,\alpha_i)}{\partial\theta}, \qquad \ell_{\alpha_i} = \sum_{t=1}^{T} \frac{\partial \log f(z_{it}|\theta,\alpha_i)}{\partial\alpha_i},$$

$$\ell_{\theta\theta} = \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{\partial^2 \log f(z_{it}|\theta,\alpha_i)}{\partial\theta\partial\theta'}, \qquad \ell_{\alpha_i\alpha_i} = \sum_{t=1}^{T} \frac{\partial^2 \log f(z_{it}|\theta,\alpha_i)}{\partial\alpha_i\partial\alpha_i'},$$

and

$$\ell_{\theta\alpha_i} = \sum_{t=1}^{T} \frac{\partial^2 \log f(z_{it}|\theta,\alpha_i)}{\partial\theta\partial\alpha_i'} = \ell'_{\alpha_i\theta}.$$

By making use of partitioned-invere formulae we arrive at an expression for the inverse Hessian, ℓ^{-1}, that can be computed by using only the inverses of the substantially smaller matrices $\ell_{\theta\theta}$ and $\ell_{\alpha_i\alpha_i}$.

A Newton step for θ then is

$$\theta - (\ell^{-1})_{\theta\theta}\,\ell_\theta - \sum_{i=1}^{N}(\ell^{-1})_{\theta\alpha_i}\,\ell_{\alpha_i} = \theta - (\ell^{-1})_{\theta\theta}\left(\ell_\theta - \sum_{i=1}^{N}\ell_{\theta\alpha_i}\,\ell^{-1}_{\alpha_i\alpha_i}\,\ell_{\alpha_i}\right).$$

A Newton step for α_i then is

$$\begin{aligned}&\alpha_i - (\ell^{-1})_{\alpha_i\theta}\,\ell_\theta - \sum_{j=1}^{N}(\ell^{-1})_{\alpha_i\alpha_j}\,\ell_{\alpha_j}\\ =&\alpha_i - \ell^{-1}_{\alpha_i\alpha_i}\left(\ell_{\alpha_i} - \ell_{\alpha_i\theta}(\ell^{-1})_{\theta\theta}\left(\ell_\theta - \sum_{j=1}^{N}\ell_{\theta\alpha_j}\,\ell^{-1}_{\alpha_j\alpha_j}\,\ell_{\alpha_j}\right)\right).\end{aligned}$$

Outline

Type heterogeneity

An alternative to fixed-effect and random-effect specifications is to consider a finite-mixture model.

Here, the distribution of α_i is not specified but its support is known to consist of a fixed number of m points.

Conditional-independence restrictions yield (nonparametric) identification from short panel data provided the type-specific distributions are linearly independent.

We give a simple proof below that assumes stationarity.

In it we let

$$G_z(y) := \mathbb{P}(y_{it} \leq y | \alpha = z).$$

Then identification follows from linear independence of $G_1, \ldots G_m$ if $T \geq 3$.

Identification

Consider a grid of m points $y_1, \dots y_m$ and construct the $m \times m$ matrix

$$\boldsymbol{G} = \begin{pmatrix} G_1(y_1) & \dots & G_m(y_1) \\ \vdots & \dots & \vdots \\ G_1(y_m) & \dots & G_m(y_m) \end{pmatrix}.$$

By linear independence we can always find points such that this matrix has maximal column rank.

Next, let

$$(\boldsymbol{a})_{m_1} = \mathbb{P}(y_{i1} \leq y_{m_1}),$$

and

$$(\boldsymbol{A})_{m_1,m_2} = \mathbb{P}(y_{i1} \leq y_{m_1}, y_{i2} \leq y_{m_2}),$$

and

$$(\boldsymbol{A}_y)_{m_1,m_2} = \mathbb{P}(y_{i1} \leq y_{m_1}, y_{i2} \leq y_{m_2}, y_{i3} \leq y),$$

for $(m_1, m_2) \in \{1, \dots, m\}^2$ and any y.

Note that

$$\boldsymbol{a} = \boldsymbol{G}\,\boldsymbol{p},$$

for $\boldsymbol{p} := (p_1, \ldots, p_m)'$ with $p_z := \mathbb{P}(\alpha = z)$.

Similarly, by conditional independence of outcomes given types,

$$\boldsymbol{A} = \boldsymbol{G}\boldsymbol{D}\boldsymbol{G}'$$

and

$$\boldsymbol{A}_y = \boldsymbol{G}\boldsymbol{D}^{1/2}\boldsymbol{D}_y\boldsymbol{D}^{1/2}\boldsymbol{G}'$$

for $\boldsymbol{D} := \mathrm{diag}(p_1, \ldots, p_m)$ and $\boldsymbol{D}_y = \mathrm{diag}(G_1(y), \ldots, G_m(y))$ for any y.

Next, we first recover $\boldsymbol{D}_y$ for any y, yielding the conditional distributions of outcomes given types. Given these, we may then recover the distribution of types.

All of this is up to an arbitrary permutation of the types.

Use an eigendecomposition of $\boldsymbol{A}$ to construct a matrix $\boldsymbol{V}$ so that

$$\boldsymbol{V}\boldsymbol{A}\boldsymbol{V}^\top = \boldsymbol{I}_m.$$

Note that

$$\boldsymbol{V}\boldsymbol{A}\boldsymbol{V}^\top = \boldsymbol{V}(\boldsymbol{G}\boldsymbol{D}\boldsymbol{G}')\boldsymbol{V}^\top = (\boldsymbol{V}\boldsymbol{G}\boldsymbol{D}^{1/2})(\boldsymbol{D}^{1/2}\boldsymbol{G}'\boldsymbol{V}^\top) = \boldsymbol{Q}\,\boldsymbol{Q}' = \boldsymbol{I}_m$$

for $\boldsymbol{Q} := \boldsymbol{V}\boldsymbol{G}\boldsymbol{D}^{1/2}$ an $m \times m$ orthonormal matrix of full rank.

Next,

$$\boldsymbol{V}\boldsymbol{A}_y\boldsymbol{V}^\top = \boldsymbol{V}(\boldsymbol{G}\boldsymbol{D}^{1/2}\boldsymbol{D}_y\boldsymbol{D}^{1/2}\boldsymbol{G}')\boldsymbol{V}^\top = (\boldsymbol{V}\boldsymbol{G}\boldsymbol{D}^{1/2})\boldsymbol{D}_y(\boldsymbol{D}^{1/2}\boldsymbol{G}'\boldsymbol{V}^\top) = \boldsymbol{Q}\boldsymbol{D}_y\boldsymbol{Q}'$$

for any y.

By linear independence of the columns of $\boldsymbol{G}$ we can recover the matrix $\boldsymbol{Q}$ up to sign and permutation of its columns as the joint diagonalizer of the collection of matrices $\boldsymbol{V}\boldsymbol{A}_{y_1}\boldsymbol{V}^\top, \ldots, \boldsymbol{V}\boldsymbol{A}_{y_m}\boldsymbol{V}^\top$. Let this matrix be denoted as $\bar{\boldsymbol{Q}} = \boldsymbol{Q}\boldsymbol{\Delta}\boldsymbol{\Sigma}$.

Here, $\boldsymbol{\Sigma}$ is a permutation matrix and $\boldsymbol{\Delta}$ is a diagonal matrix with only 1 or -1 as diagonal entries.

With $\bar{\boldsymbol{Q}}$ in hand,

$$\bar{\boldsymbol{D}}_y = \bar{\boldsymbol{Q}}'\boldsymbol{V}\boldsymbol{A}_y\boldsymbol{V}'\bar{\boldsymbol{Q}} = \boldsymbol{\Sigma}'\boldsymbol{\Delta}\boldsymbol{Q}'\boldsymbol{V}\boldsymbol{A}_y\boldsymbol{V}'\boldsymbol{Q}\boldsymbol{\Delta}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}'\boldsymbol{D}_y\boldsymbol{\Sigma}$$

gives the conditional distributions $G_1, \ldots, G_m$ at point y, for some ordering of the latent types.

Given these, we know $\boldsymbol{G}$ up to the same ordering of its columns, i.e., we know $\bar{\boldsymbol{G}} := \boldsymbol{G}\boldsymbol{\Sigma}$, and so

$$\boldsymbol{a} = \boldsymbol{G}\boldsymbol{p} = \boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{\Sigma}'\boldsymbol{p} = \bar{\boldsymbol{G}}\bar{\boldsymbol{p}}$$

yields

$$\bar{\boldsymbol{p}} = (\bar{\boldsymbol{G}}'\bar{\boldsymbol{G}})^{-1}\bar{\boldsymbol{G}}'\boldsymbol{a},$$

i.e., the type distribution up to the same permutation of types as before.

This constructive result has been used to build a nonparametric estimator. (No discussion on this here.)

Parametric estimation

Take a parametric model without covariates for simplicity of notation.

A unit i of type z has (conditional) likelihood

$$\ell_z(y_{i1}, \ldots, y_{iT}; \theta_z)$$

and marginal likelihood

$$\sum_{z=1}^{m} \omega_z \, \ell_z(y_{i1}, \ldots, y_{iT}; \theta_z), \qquad \omega_z = \mathbb{P}(\alpha_i = z).$$

Classification

We do not aim to estimate the type, but rather the conditional distribution of the data given types, $\ell_1, \ldots, \ell_m$, as well as the type distribution, $\omega_1, \ldots, \omega_m$.

From Bayes'rule,

$$\mathbb{P}(\alpha = z | y_{i1}, \ldots, y_{iT}) = \frac{\omega_z \, \ell_z(y_{i1}, \ldots, y_{iT}; \theta_z)}{\sum_{z'=1}^{m} \omega_{z'} \, \ell_{z'}(y_{i1}, \ldots, y_{iT}; \theta_{z'})}$$

is then equally known.

This allows to perform posterior classification.

EM algorithm

The log-likelihood for a random sample is

$$\sum_{i=1}^{N} \log \left(\sum_{z} \omega_z \, \ell_z(y_{i1}, \ldots, y_{iT}; \theta_z) \right)$$

which is typically difficult to optimize.

Consider data augmentation. If we would know α_i then we would consider the complete-data problem.

Here, the log-likelihood is

$$\sum_{i=1}^{N} \sum_{z} \{\alpha_i = z\} (\log \omega_z + \log \ell_z(y_{i1}, \ldots, y_{iT}, \theta_z)).$$

Collect all the parameters $\theta_1, \ldots, \theta_m$ and $\omega_1, \ldots, \omega_m$ in the vector ϑ.

Let ϑ_1 be an initial guess.

Then EM updates to ϑ_2 via the following two-step routine.

1. Compute the expected log-likelihood conditional on the data using ϑ_1:

$$\sum_{i=1}^{N} \sum_{z} \mathbb{E}_{\vartheta_1} (\{\alpha_i = z\} | y_{i1}, \ldots, y_{iT}) (\log \omega_z + \log \ell_z(y_{i1}, \ldots, y_{iT}, \theta_z)) .$$

Note that this amounts to weighting by a (probabilistic) classification of units to types.

2. Maximize this expected log-likelihood to find ϑ_2. This is usually much easier than maximizing the likelihood itself, and is often feasible in closed form.