INFERENCE IN PANEL DATA AND NETWORK DATA

(M2 ETE Econometrics II)

Koen Jochmans

Toulouse School of Economics

Last revised on August 2, 2026

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 $N >> T$.

Often 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.

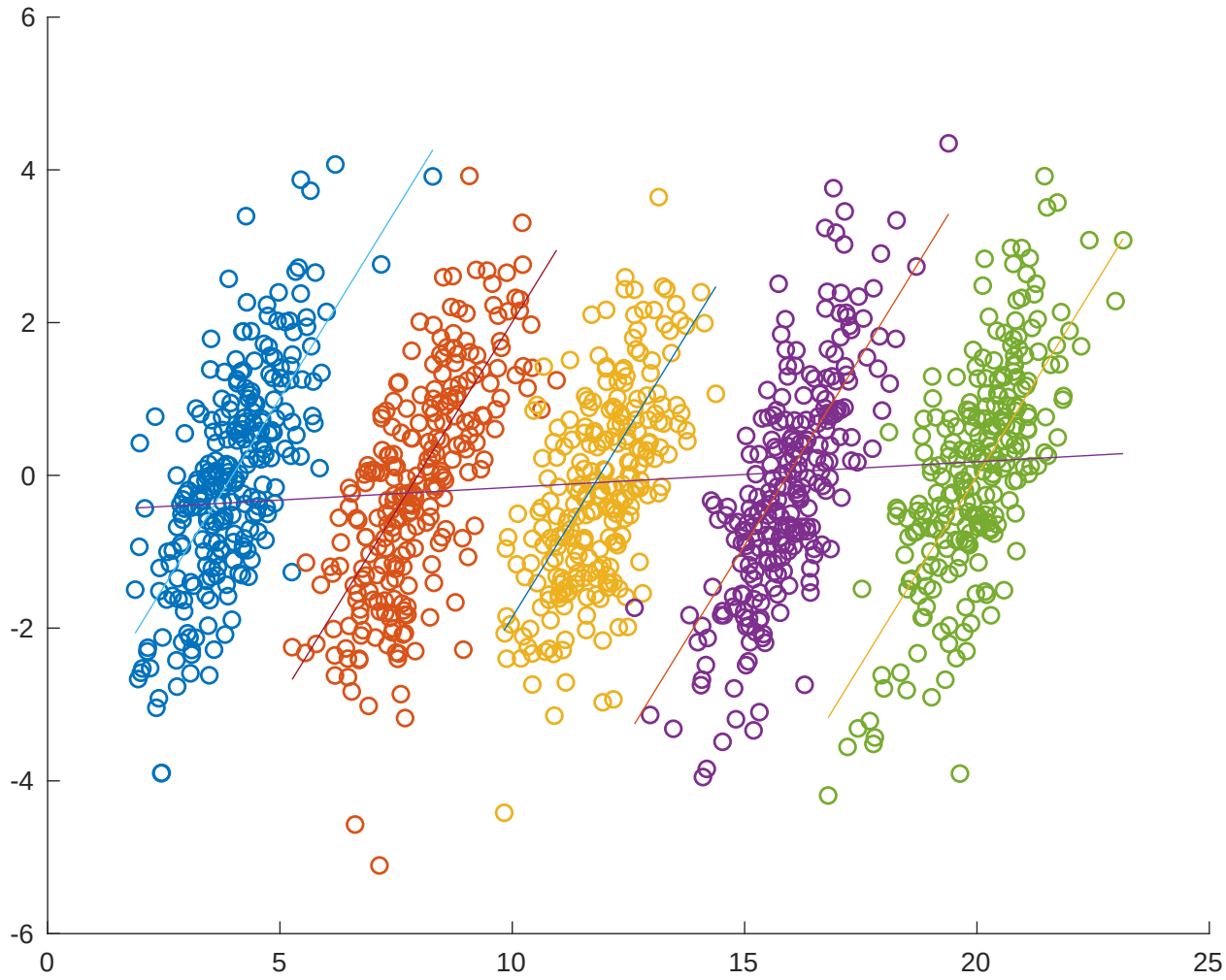
- y_{it} is output;
- x_{it} are observable inputs ;
- α_i is soil quality;
- ε_{it} is rainfall (unpredictable).

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 as regressor.)

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

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

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,

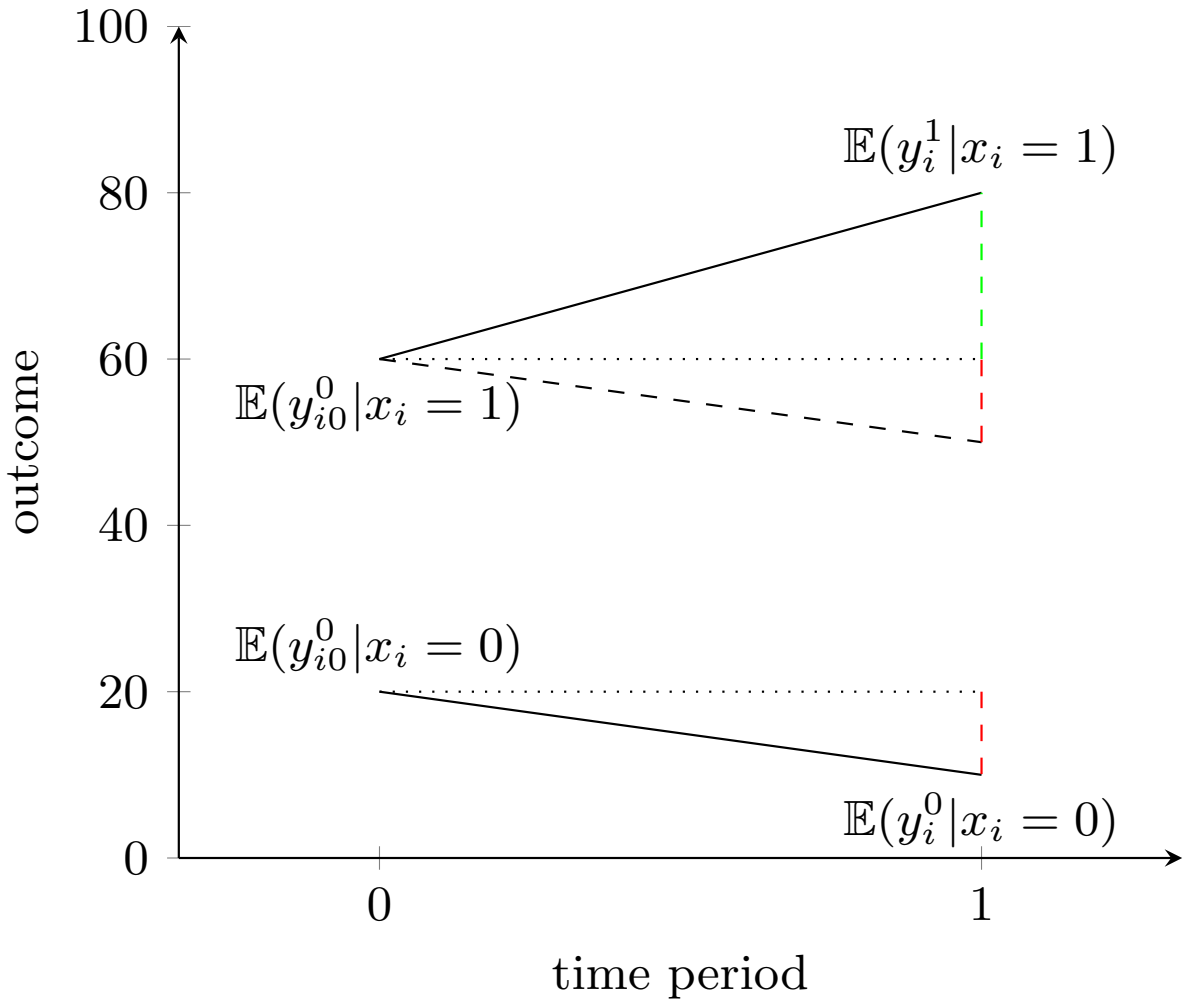
$$\begin{aligned}
\mathbb{E}(y_i - y_{i0}|x_i = 1) =& \mathbb{E}(y_i^0 - y_{i0}^0|x_i = 1) + \mathbb{E}(y_i^1 - y_i^0|x_i = 1), \\
\mathbb{E}(y_i - y_{i0}|x_i = 0) =& \mathbb{E}(y_i^0 - y_{i0}^0|x_i = 0),
\end{aligned}$$

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:

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

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

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

and ι_T a vector of ones.

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:

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

Then

$$\mathbb{E}(Dy_i - DX_i\beta | X_i) = 0$$

holds, yielding unconditional moment conditions.

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, so FD least-squares would be inefficient.

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 $B = I_N \otimes \iota_T$ and $\alpha = (\alpha_1, \dots, \alpha_N)'$.

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,

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

so that $B(B'B)^{-1}B' = (I_N \otimes \iota_T)(I_N \otimes {}^1\!/_T)(I_N \otimes \iota_T)' = (I_N \otimes {}^{\iota_T\iota_T'}\!/_T)$, 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. Here, we estimate (the realizations of) $\alpha_1, \ldots, \alpha_N$ as unit-specific intercepts.

WG via a control-function regression

Consider pooling the data and regressing y_{it} on a (common) constant, x_{it} and the group-specific mean $\bar{x}_i$.

(taking x_{it} to be scalar for simplicity,) by partitioned regression this amounts to regressing y_{it} on the residual from a regression of x_{it} on a constant and $\bar{x}_i$, say, $\hat{v}_{it} = x_{it} - \hat{\delta}_1 - \hat{\delta}_2 \bar{x}_i$ for

$$(\hat{\delta}_1, \hat{\delta}_2) = \arg\min_{\delta_1, \delta_2} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \delta_1 - \delta_2 \bar{x}_i)^2.$$

Applying 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 $\hat{v}_{it} = x_{it} - \bar{x}_i$ and we estimate β as

$$\frac{\sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_i) y_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_i) x_{it}} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{\sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)^2} = \hat{\beta}_{\text{WG}}.$$

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 would usually not be willing to 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, as $N \to \infty$ with T held fixed,

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

The inconsistency of $\hat{\sigma}^2$ is a consequence of the demeaning of the data or, equivalently, due to the estimation noise in the fixed effects (the need to estimate group-specific intercepts).

Here, a simple 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 in contrast 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.

To is so because

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

This is essentially the same problem as in the homoskedastic case.

As $N\to\infty$ but T remains fixed we are working with $\varepsilon_{it}-\bar{\varepsilon}_i$, not with ε_{it}.

Comment: Specification testing

The above also highlights that all conventional 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

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

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 there is no endogeneity problem and a simple pooled least-squares regression is consistent.

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

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 (and illuminating) 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, given the covariance structure of $\varepsilon_{it} + \eta_i$.

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,

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

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

To illustrate, suppose we care about

$$\theta = \mathbb{E}(\varphi(z_{it}; \alpha_i))$$

for some known function φ.

This could be a marginal effect or an estimating equation, for example.

We estimate θ by

$$\hat{\theta} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \varphi(z_{it}, \hat{\alpha}_i).$$

This estimator depends on noisy estimates of the α_i, which introduce bias.

We can Taylor expand φ to second order and take expectations to arrive at

$$\mathbb{E}(\hat{\theta} - \theta) = \frac{\sum_{i=1}^{N} \operatorname{cov}(\varphi'(z_{it}, \alpha_i), \hat{\alpha}_i - \alpha_i) + \mathbb{E}(\varphi'(z_{it}, \alpha_i)) \, \mathbb{E}(\hat{\alpha}_i - \alpha_i)}{N}$$
$$+ \frac{\sum_{i=1}^{N} 1/2 \, \mathbb{E}(\varphi''(z_{it}, \alpha_i)) \, \mathbb{E}((\hat{\alpha}_i - \alpha_i)^2)}{N} + O(T^{-2})$$

The (leading) bias has three components, each coming from a different source:

-correlation between the data z_{it} and the estimator $\hat{\alpha}_i$,

-bias in the estimator $\hat{\alpha}_i$,

-variance in the estimator $\hat{\alpha}_i$.

In general, each of these is of order T^{-1}.

The third source implies that any nonlinear function of an unbiased estimator is biased.

Any nonlinear estimating equation will suffer from bias unless something 'miraculous' happens.

Under the classical large-N, fixed-T asymptotic framework, $O(T^{-1})$ bias does not vanish, leading to inconsistency.

Linking this to linear models

In

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

we estimate α_i (given β) as

$$\hat{\alpha}_i(\beta) = \bar{y}_i - \bar{x}_i'\beta = \alpha_i + \bar{\varepsilon}_i,$$

which is unbiased because

$$\mathbb{E}(\hat{\alpha}_i(\beta)|x_{i1}, \ldots, x_{iT}, \alpha_i) = \alpha_i = \mathbb{E}(y_{it} - x_{it}'\beta|x_{i1}, \ldots, x_{iT}, \alpha_i)$$

The normal equation for β is

$$\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}(y_{it} - \hat{\alpha}_i(\beta) - x_{it}'\beta),$$

which has mean zero by the unbiasedness from above, as can be seen by iterating expectations.

Nevertheless, if we care about, say, the second moment of the fixed effects, $1/N \sum_{i=1}^{N} \alpha_i^2$, this is biased even if β would be known. In that case, we exactly have $\hat{\alpha}_i = \alpha_i + \bar{\varepsilon}_i$, and so

$$\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\alpha}_i^2\right) = \frac{1}{N}\sum_{i=1}^{N}\alpha_i^2 + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\bar{\varepsilon}_i^2) \neq \frac{1}{N}\sum_{i=1}^{N}\alpha_i^2$$

(where we are treating the α_i as fixed in this particular case).

In the (first-order) autoregressive problem, where $x_{it} = y_{it-1}$, we still have

$$\hat{\alpha}_i(\beta) = \bar{y}_i - \bar{x}_i'\beta = \alpha_i + \bar{\varepsilon}_i,$$

but now

$$\mathbb{E}(x_{it}\hat{\alpha}_i(\beta)) = \mathbb{E}(y_{it-1}\alpha_i) + \mathbb{E}(y_{it-1}\bar{\varepsilon}_i) \neq \mathbb{E}(y_{it-1}\alpha_i) = \mathbb{E}(y_{it-1}(y_{it} - y_{it-1}\beta)).$$

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

$$y_{it} = \max\{\alpha_i + x_{it}\beta + \varepsilon_{it}, 0\}$$

with stationary errors that are independent of regressors.

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,

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

equals

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

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,

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

is zero by stationarity.

The errors are unobserved but

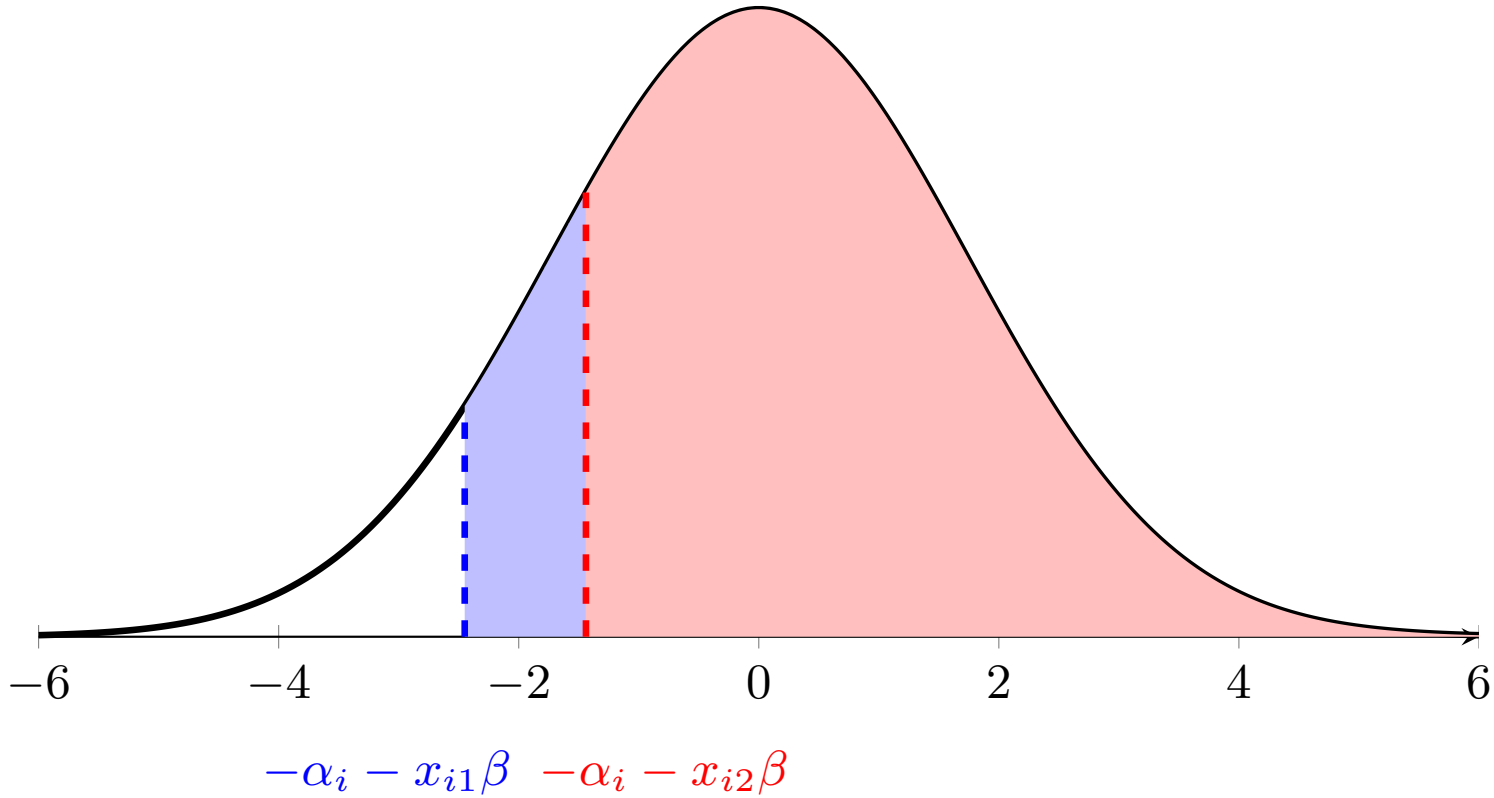
$$\begin{aligned}\varepsilon_{i1} > -\alpha_i - x_{i1}\beta &\Leftrightarrow y_{i1} > 0\\ \varepsilon_{i1} > -\alpha_i - x_{i2}\beta &\Leftrightarrow y_{i1} > -\Delta x_i\beta,\end{aligned}$$

and, similarly,

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

We thus consider a trimmed least-squares estimator

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


−6
−4
−2
0
2
4
6
$-\alpha_i - x_{i1}\beta$
$-\alpha_i - x_{i2}\beta$

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, 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 $N/T \to c^2 \in [0, \infty)$ 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 $N/T \to c^2$,

Fixed-effect estimator satisfies

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

(for Σ some asymptotic variance).

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

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

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.

Bootstrap inference

An alternative to recentering the limit distribution by bias correction is to rely on the bootstrap.

Let $\hat{\theta}^*$ denote an estimator computed from a bootstrap sample.

Then quite generally a bootstrap scheme can be concocted so that

$$\sqrt{NT}(\hat{\theta}^* - \hat{\theta}) \xrightarrow{d^*} \mathbf{N}(cB, \Sigma),$$

as $N/T \to c^2 \in [0.\infty)$.

That is, the bootstrap mimics the biased fixed-effect estimator, which had

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

The implication is that one can do inference using the bootstrap, ignoring the incidental-parameter bias.

To illustrate, let

$$\hat{Q}_\tau = \inf\{\hat{Q} : \tau \leq \mathbb{P}^*(r'(\hat{\theta}^* - \hat{\theta}) \leq \hat{Q})\},$$

the quantile distribution of the bootstrap distribution.

Then $\hat{Q}_\tau$ estimates

$$Q_\tau = \inf\{Q : \tau \leq \mathbb{P}(r'(\hat{\theta} - \theta) \leq Q)\},$$

the quantile distribution of the actual distribution.

Then the set

$$\{r'\vartheta : r'\hat{\theta} - \hat{Q}_\tau \leq r'\vartheta\}$$

is an upper one-sided confidence interval with asymptotic coverage τ.

This implementation is the same as if there was no bias.

Can also use the bootstrap distribution for critical values for test statistics.

Likelihood models

In the likelihood case we have density/mass function of the form

$$f(y_{it}|y_{it-1},\ldots,y_{it-p},x_{it};\theta,\alpha_i).$$

Can compute the maximum-likelihood estimator $\hat{\theta}$ and $\hat{\alpha}_1,\ldots,\hat{\alpha}_N$ to fit the transition kernel and generate bootstrap data from

$$f(y^*_{it}|y^*_{it-1},\ldots,y^*_{it-p},x^*_{it};\hat{\theta},\hat{\alpha}_i),$$

where we set $x^*_{it} = x_{it}$ and initialize the process at the observed $y_{i0},\ldots y_{it-p+1}$.

To see why this works, take again

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

As before, $\hat{\alpha}_i = \bar{y}_i$,

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

which has bias $-\theta/T$ because $\mathbb{E}((y_{it} - \bar{y}_i)^2) = \theta - \theta/T$.

Simulate bootstrap data from $y^*_{it} \sim \mathbf{N}(\hat{\alpha}_i, \hat{\theta})$ to get the bootstrap estimator

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

Now, $\mathbb{E}^*((y^*_{it} - \bar{y}^*_i)^2) = \hat{\theta} - \hat{\theta}/T$ in the same way as before. But then we have

$$\mathbb{E}^*(\hat{\theta}^* - \hat{\theta}) = -\frac{\hat{\theta}}{T} = -\frac{\theta}{T} + \frac{\theta}{T^2} + O_p\left(\frac{1}{\sqrt{NT^3}}\right),$$

which is first-order correct.

Other bootstrap schemes

Not all bootstrap schemes mimic incidental-parameter bias!

For example, cross-sectional resampling of entire time series will not work here.

What does work (in addition to the parametric bootstrap) is as follows:

In models with additive errors we can rely on a residual bootstrap.

In nonlinear models we can resample entire cross-sections.

For the latter, dynamics can be handled by using a moving-block version of Efron's bootstrap.

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 of the form

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

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-inverse 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

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

Numerical example

Binary-choice autoregressive logit

$$y_{it} = \begin{cases} 1 & \text{if } \alpha_i + \theta y_{it-1} > \varepsilon_{it} \\ 0 & \text{if not} \end{cases},$$

with $\alpha_i = 0$ and steady-state initial conditions y_{i0}.

			BIAS	STD	COVERAGE							SIZE		
θ	N	T	MLE		MLE	BC1	BC2	BB	DBB	SB	DSB	LR	LR*	LR**
1/2	100	10	-0.461	0.145	0.111	0.942	0.970	0.964	0.958	0.928	0.940	0.887	0.064	0.041
1/2	100	20	-0.219	0.095	0.378	0.952	0.962	0.958	0.955	0.943	0.949	0.640	0.057	0.048
1/2	250	10	-0.459	0.090	0.001	0.895	0.968	0.957	0.949	0.928	0.932	0.999	0.092	0.051
1/2	250	20	-0.217	0.062	0.054	0.937	0.952	0.958	0.961	0.942	0.945	0.957	0.062	0.049
1	100	10	-0.514	0.150	0.086	0.880	0.941	0.964	0.949	0.916	0.937	0.919	0.113	0.065
1	100	20	-0.244	0.103	0.332	0.907	0.921	0.948	0.948	0.931	0.940	0.627	0.055	0.040
1	250	10	-0.513	0.095	0.000	0.745	0.898	0.970	0.952	0.907	0.952	0.999	0.144	0.053
1	250	20	-0.244	0.065	0.039	0.881	0.922	0.959	0.950	0.954	0.944	0.957	0.069	0.053
3/2	100	10	-0.623	0.165	0.034	0.695	0.837	0.942	0.930	0.867	0.926	0.966	0.178	0.091
3/2	100	20	-0.299	0.113	0.269	0.835	0.883	0.940	0.930	0.932	0.932	0.712	0.068	0.047
3/2	250	10	-0.624	0.104	0.000	0.304	0.593	0.917	0.913	0.776	0.936	1.000	0.274	0.109
3/2	250	20	-0.302	0.071	0.010	0.636	0.741	0.953	0.949	0.932	0.948	0.981	0.101	0.057

BC1 and BC2 are analytical bias corrections. BB and SB are the bootstrap. DBB and DSB are double-bootstrap implementations.

Higher-order bias correction

In smooth cases,

$$\text{plim}_{N\to\infty}\hat{\theta} = \theta + \frac{B_1}{T} + \frac{B_2}{T^2} + \cdots + \frac{B_q}{T^q} + O(T^{-(q+1)}).$$

A first-order correction such as the jackknife than has bias of order T^{-2}.

Thus, higher-order bias remains asymptotically important unless $N/T^3 \to 0$.

Can then perform higher-order bias correction to remove additional bias terms:

Correction to order q has remaining bias of order $T^{-(q+1)}$, and so will be correctly centered as long as $N/T^{2q+1} \to 0$.

Analytical corrections of this kind are difficult, but a jackknife can again be applied.

Consider a correction of the form

$$\check{\theta} = \hat{\theta} - \frac{\hat{B}_1}{T}.$$

Then $\hat{B}_1$ itself will itself have a bias of order T^{-1}; say

$$\text{plim}_{N\to\infty}\hat{B}_1 = B_1 + \frac{C_1}{T} + O(T^{-2}).$$

This implies that

$$\text{plim}_{N\to\infty}\check{\theta} = \theta + \frac{B_2 - C_1}{T^2} + O(T^{-3}).$$

A first-order bias correction affects the remaining higher-order bias at all orders.

To illustrate the jackknife it suffices to consider a second-order adjustment.

Suppose that T is divisible by both 2 and 3. Consider an estimator of the form

$$(1 + a_{1/2} + a_{1/3})\hat{\theta} - a_{1/2}\bar{\theta}_{1/2} - a_{1/3}\bar{\theta}_{1/3},$$

which is some linear combination of the original estimator and (averages of) estimators from subpanels of length $T/2$ and $T/3$.

Then we need to find coefficients $a_{1/2}$ and $a_{1/3}$ so that

$$\left(\frac{1 + a_{1/2} + a_{1/3}}{T} - 2\frac{a_{1/2}}{T} - 3\frac{a_{1/3}}{T}\right) B_1 = 0,$$
$$\left(\frac{1 + a_{1/2} + a_{1/3}}{T^2} - 2^2\frac{a_{1/2}}{T^2} - 3^2\frac{a_{1/3}}{T^2}\right) B_2 = 0,$$

for any (B_1, B_2).

This has solution $a_{1/2} = 3$, $a_{1/3} = -1$, and so a second-order corrected jackknife is

$$3\hat{\theta} - 3\bar{\theta}_{1/2} + \bar{\theta}_{1/3}.$$

This remains correctly centered as long as $N/T^5 \to 0$.

Iterating the bootstrap

A similar improvement can equally be obtained through use of the bootstrap applied to the uncorrected estimator.

Here, this works by iterating the bootstrap.

Write

$$S = \sqrt{NT}(\hat{\theta} - \theta) \xrightarrow{d} B + Z, \qquad S^* = \sqrt{NT}(\hat{\theta}^* - \hat{\theta}) \xrightarrow{d^*} \hat{B} + Z^*,$$

where B and $\hat{B}$ are the respective (scaled) bias terms and Z and Z^* are mean zero normally-distributed random variables.

Focus on the (left-sided) bootstrap p-value,

$$\check{P} = \mathbb{P}^*(S^* \leq S) \xrightarrow{p} \mathbb{P}^*(Z^* \leq Z - (\hat{B} - B)) \xrightarrow{p} V + O_p(\sqrt{N/T^3})$$

for $V \sim \text{Uniform}[0, 1]$.

The intuition is again that the bias is only correctly replicated to first order.

Now consider a second-layer bootstrap, where

$$S^{**} = \sqrt{NT}(\hat{\theta}^{**} - \hat{\theta}^{*}) \xrightarrow{d^{**}} \hat{B}^{*} + Z^{**}.$$

Then $\hat{B}^{*}$ is a biased estimator of $\hat{B}$ in the same way that $\hat{B}$ is for B. Moreover,

$$(\hat{B}^{*} - \hat{B}) = (\hat{B} - B) + O_p(\sqrt{N/T^5})$$

Now consider

$$\check{\check{P}} = \mathbb{P}^{*}(\check{P}^{*} \leq \check{P}), \qquad \check{P}^{*} = \mathbb{P}^{**}(S^{**} \leq S^{*}).$$

Then

$$\check{\check{P}} \xrightarrow{p} \mathbb{P}^{*}(Z^{*} - (\hat{B}^{*} - \hat{B}) \leq Z - (\hat{B} - B)) \to V + O_p(\sqrt{N/T^5})$$

The errors induced by the (first layer) bootstrap at order two is correctly replicated by the second layer bootstrap. This explains why an improvement is achieved.

The autoregressive model

Look at

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

where within-groups is subject the Nickell bias.

Let

$$\tau^2 = \frac{\mathbb{E}((y_{i0} - \mu_i)^2)}{\gamma^2}, \qquad \mu_i = \frac{\alpha_i}{1-\theta}, \qquad \gamma^2 = \frac{\sigma^2}{1-\theta^2}.$$

Then the bias expansion to order two for $\hat{\theta}$ is

$$\frac{c_1(\theta)}{T} + \frac{c_2(\theta, \tau^2)}{T^2} + O(T^{-3}),$$

for

$$c_1(\theta) = -(1+\theta), \qquad c_2(\theta, \tau^2) = -(1-\theta)^{-1}\left(\theta(1+\theta) - (\tau^2 - 1)\right).$$

The second (and higher) order bias depends on the initial conditions through their deviation from steady state.

We simulate the distribution of p-values for

$$T = 2\lceil\sqrt[3]{N}\rceil, \qquad \text{and} \qquad N \in \{100, 250, 500, 1000, 5000, 10000\}.$$

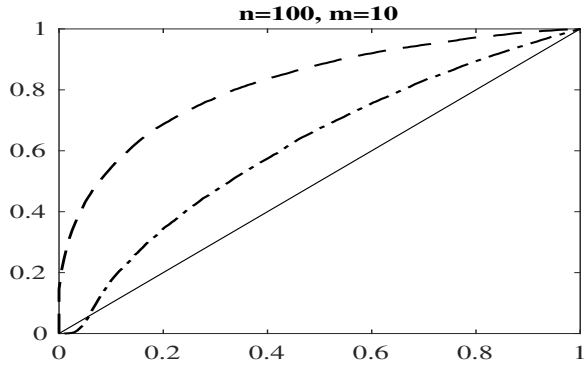
This mimics an asymptotic where $N/T^3 \to 1/2$.

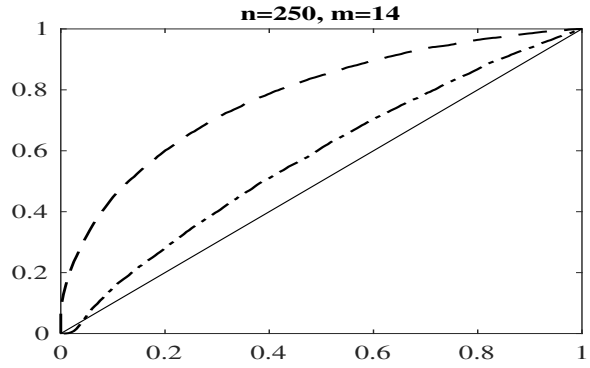
Reference is 45° line.

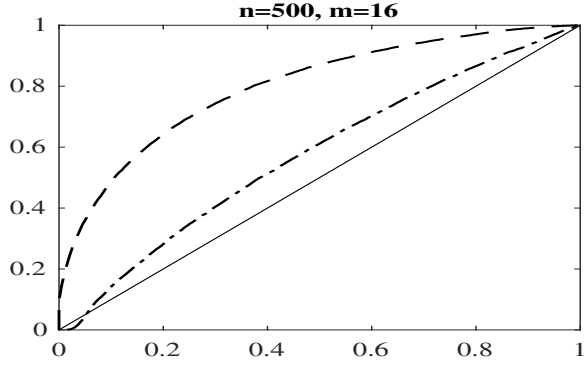
Model parameters are $\theta = 1/2$ and $\tau^2 = 0$ (by setting all α_i and y_{i0} to zero).

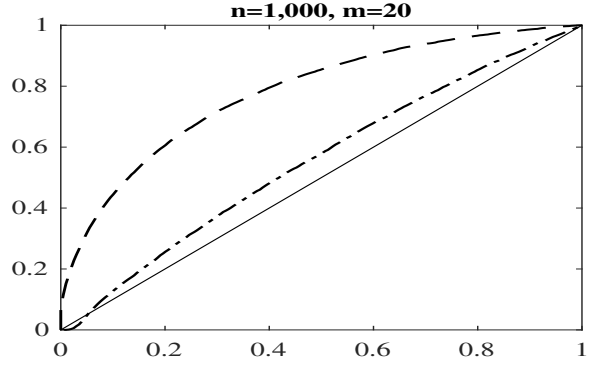
Standard-normal innovations.

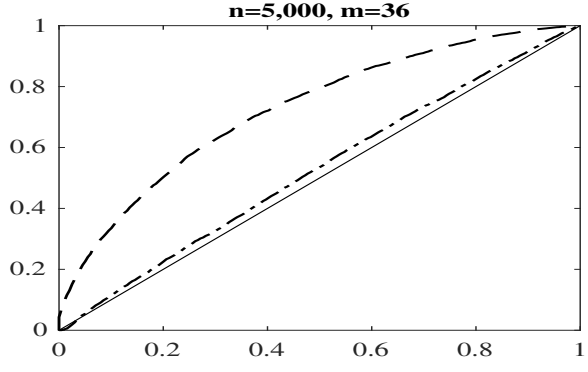
Residual bootstrap (by unit).

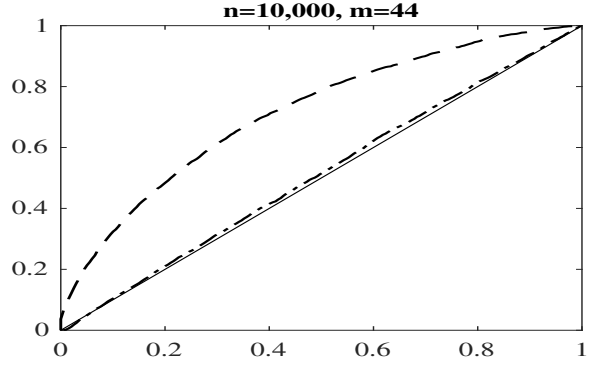

n=100, m=10
0
0.2
0.4
0.6
0.8
1


n=250, m=14
0
0.2
0.4
0.6
0.8
1


n=500, m=16
0
0.2
0.4
0.6
0.8
1


n=1,000, m=20
0
0.2
0.4
0.6
0.8
1


n=5,000, m=36
0
0.2
0.4
0.6
0.8
1


n=10,000, m=44
0
0.2
0.4
0.6
0.8
1

Outline

Type heterogeneity

An alternative modelling approach is one of type heterogeneity.

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

The discreteness helps both in a fixed-effect and random-effect framework.

Working with types is common in structural econometric models but has been gaining popularity elsewhere.

Here we consider a linear regression model with group fixed effects, and nonparametric finite-mixture models.

Consider again

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

but now suppose that α_i can take on only m values. This means that the population is made up of m clusters/groups, $c = 1, \ldots, m$.

If we would know cluster membership, we would observe the variable z_i

$$z_i = \begin{cases} 1 & \text{if } i \text{ is in cluster } 1 \\ \vdots & \vdots \\ m & \text{if } i \text{ is in cluster } m \end{cases},$$

and could run a simple pooled regression of y_{it} on x_{it} and the m dummy variables

$$g_{ic} = \begin{cases} 1 & \text{if } i \text{ is in cluster } c \\ 0 & \text{if not} \end{cases}.$$

This is a low-dimensional problem. We can thus estimate the fixed effects consistently as long as either (i) the panel becomes longer and/or (ii) all the clusters become large.

Infeasible regression problem

The oracle problem is thus

$$\min_{\alpha_1,\ldots,\alpha_m,\beta} \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - g_{i1}\alpha_1 - \cdots g_{im}\alpha_m - x_{it}\beta)^2.$$

This is really a cross-sectional problem; the panel structure is not important here.

We can re-write the objective as

$$\min_{\alpha_1,\ldots,\alpha_m,\beta} \sum_{z=1}^{m}\sum_{i:z_i=z}\sum_{t=1}^{T}(y_{it} - \alpha_z - x_{it}\beta)^2,$$

highlighting that this problem factors across clusters.

The difficulty in practice is that group membership is not observed.

Then we need to minimize over the α_z and the group assignment $\{z_i\}$ itself.

Classification

Fix β and write $v_{it} = y_{it} - x_{it}\beta$ for the implied error term.

Note that

$$\sum_{z=1}^{m} \sum_{i:z_i=z} \sum_{t=1}^{T} (v_{it} - \alpha_z)^2 = T \sum_{z=1}^{m} \sum_{i:z_i=z} (\bar{v}_i - \alpha_z)^2 + \sum_{i=1}^{N} \sum_{t=1}^{T} (v_{it} - \bar{v}_i)^2,$$

and only the first term depends on the classification of units to clusters.

This first term is a standard k-means classifier problem, performed on the average time-series errors.

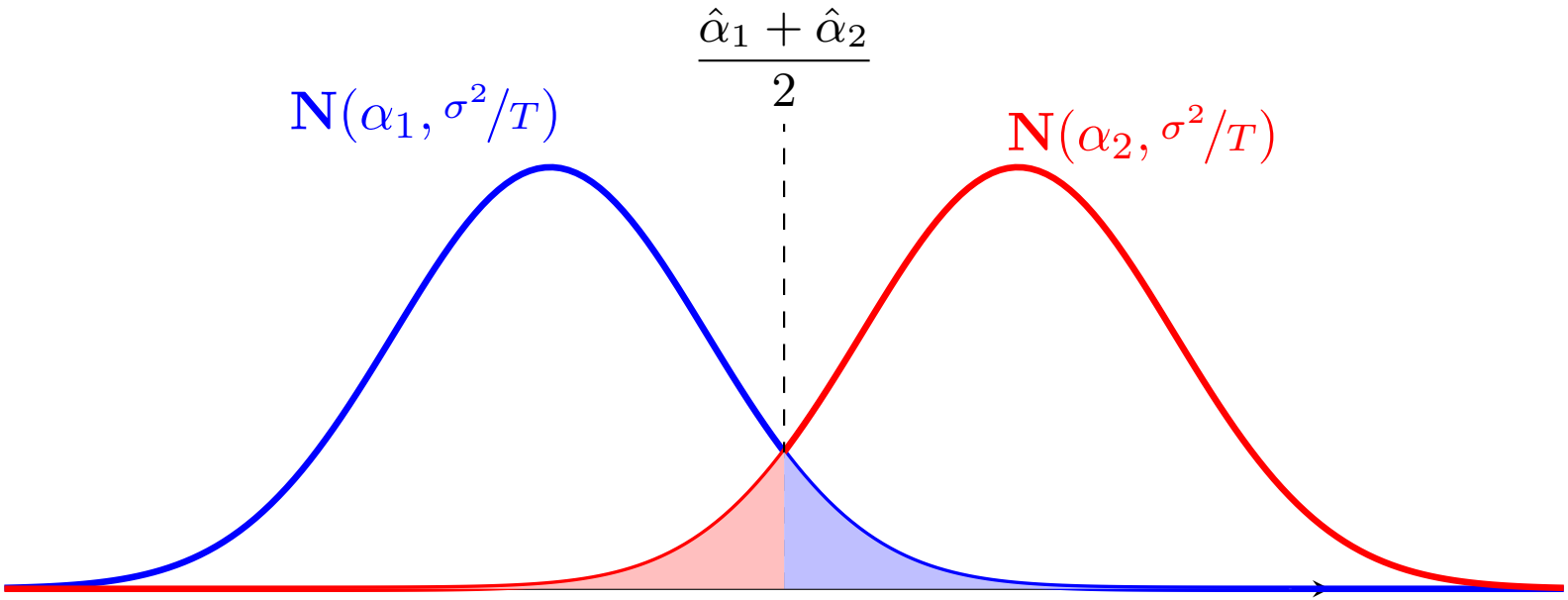
This problem corresponds to finding $m-1$ cut-off points on a line. For each unit, we choose $\hat{z}_i$ so that

$$(\bar{v}_i - \alpha_{\hat{z}_i})^2 \leq \min_z (\bar{v}_i - \alpha_z)^2.$$

This gives cut-off points between groups z and $z+1$ as

$$\frac{\hat{\alpha}_z + \hat{\alpha}_{z+1}}{2};$$

and then the estimator $\hat{\alpha}_z$ is the sample mean within group z.



The k-means approach amounts to hard classification. This requires $T \to \infty$.

The performance of this procedure depends on

- how well the distributions are separated,

- how quickly the tails vanish.

As then the misclassification can go to zero very quickly as $T \to \infty$.

An alternative is soft classification.

We do not aim to learn α_i for each i, but its (discrete) distribution.

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.

This objects are all directly observable.

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.

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.

Outline

Time effects

It is common to control for aggregate shocks through time effects.

Time dummies do not cause problems under classical asymptotics as their number is fixed.

This is not true when $T \to \infty$.

An extension of the Neyman-Scott problem is

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

Here, we have $N+T$ first-order conditions for the nuisance parameters, being

$$\sum_{t=1}^{T}(y_{it} - \hat{\alpha}_i - \hat{\gamma}_t) = 0,$$
$$\sum_{i=1}^{N}(y_{it} - \hat{\alpha}_i - \hat{\gamma}_t) = 0,$$

which is over-parametrised.

Re-arranging the normal equations gives $\hat{\alpha}_i = \bar{y}_i - \bar{\hat{\gamma}}$ and $\hat{\gamma}_t = \bar{y}_t - \bar{\hat{\alpha}}$ so that

$$\hat{\alpha}_i = (\bar{y}_i - \bar{y}) + \bar{\hat{\alpha}}, \qquad \hat{\gamma}_t = (\bar{y}_t - \bar{y}) + \bar{\hat{\gamma}},$$

and

$$\bar{y} = (\bar{\hat{\alpha}} + \bar{\hat{\gamma}}).$$

The model is invariant to a shift of the effects.

The fitted values are

$$\hat{y}_{it} = \hat{\alpha}_i + \hat{\gamma}_t = \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_t - \bar{y}).$$

The residuals are then $y_{it} - \hat{y}_{it} = (y_{it} - \bar{y}) - (\bar{y}_i - \bar{y}) - (\bar{y}_t - \bar{y})$, and the variance estimator

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

can be calculated to have mean

$$\theta \left(1 - \frac{1}{N} - \frac{1}{T} + \frac{1}{NT} \right).$$

(To see this, realize that the orthogonal projector has rank $NT - N - T + 1$.)

Limit behavior

Here, we have a limit distribution only when $N, T \to \infty$ at the same rate.

If $N/T \to c^2$,

$$\sqrt{NT}(\hat{\theta} - \theta) \xrightarrow{d} \mathbf{N}(\,(c^{-1} + c)\,\theta\,, 2\theta^2).$$

A bias correction is

$$\hat{\theta} + \frac{\hat{\theta}}{N} + \frac{\hat{\theta}}{T}.$$

The intuition of a symmetric bias decomposition extends to the more general case.

Can again jackknife, or bootstrap.

Logit

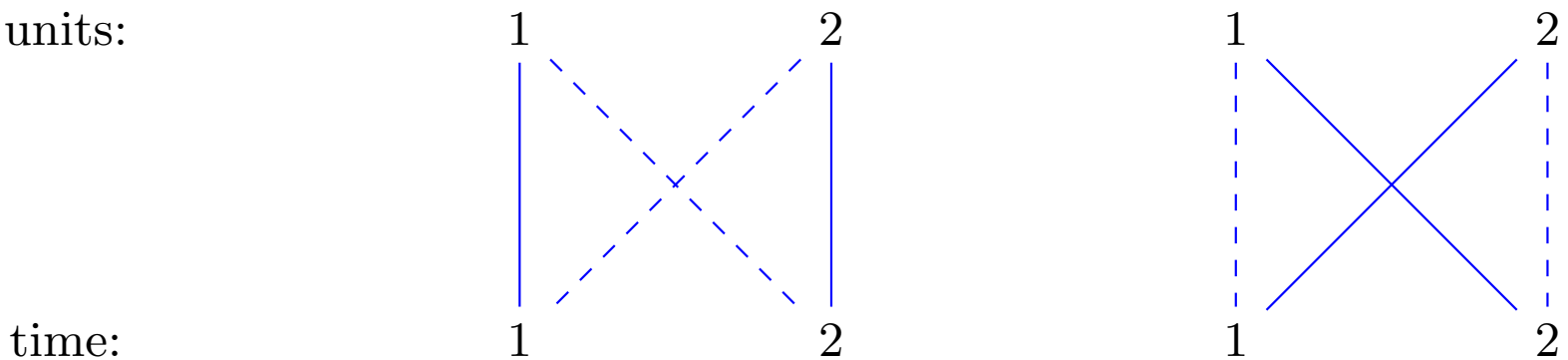
Models where an exact solution exists in the one-way case usually have a suitable extension to the two-way case.

An example is the binary-choice model

$$y_{it} = \begin{cases} 1 \text{ if } \alpha_i + \gamma_t + x_{it}\theta \geq \varepsilon_{it} \\ 0 \text{ if } \alpha_i + \gamma_t + x_{it}\theta < \varepsilon_{it} \end{cases},$$

with i.i.d. logistic errors.

Here, the building block is a pairs of units and pairs of time, with pattern



For these $z = {}^1/_2(y_{11} - y_{12}) - {}^1/_2(y_{21} - y_{22}) \in \{-1, 1\}$.

Conditional on $z \in \{-1, 1\}$ (and regressors and fixed effects), z is Bernoulli with probability

$$\frac{\mathbb{P}(z=1)}{\mathbb{P}(z=1)+\mathbb{P}(z=-1)} = \frac{1}{1+\frac{\mathbb{P}(z=-1)}{\mathbb{P}(z=\ \ 1)}}.$$

Here,

$$\frac{\mathbb{P}(z=-1)}{\mathbb{P}(z=\ \ 1)} = \frac{\mathbb{P}(y_{11}=1, y_{12}=0, y_{21}=0, y_{22}=1)}{\mathbb{P}(y_{11}=0, y_{12}=1, y_{21}=1, y_{22}=0)} = e^{-((x_{11}-x_{12})-(x_{21}-x_{22}))'\theta}$$

which does not depend on fixed effects.

The conditional success probability has a logit structure with a 'difference in differences' form.

Multiplicative-error models

The two-way version of the model with multiplicative errors is

$$y_{it} = \varphi(x_{it};\theta)\,\alpha_i\,\gamma_t\,\varepsilon_{it},$$

with errors uncorrelated over units and time.

We again work with

$$v_{it} = \frac{y_{it}}{\varphi(x_{it};\theta)},$$

which has expectation $\alpha_i\gamma_t$.

Then $v_{11}v_{22}$ and $v_{12}v_{21}$ have the same conditional mean $\alpha_1\alpha_2\gamma_1\gamma_2$ and an estimator follows from differencing.

This type of 'differencing' procedures generalize to multi-way settings, which are common in trade applications, for example. Pseudo-poisson estimators, in contrast, run into difficulties there.

Instrumental-variable versions are also available.

Outline

Social interactions

Agents do not make decisions in isolation.

- Market-entry decisions of firms.
- Effort-level decisions by workers in a team/students in a classroom.

Agents can incur indirect costs and benefits from actions of others

- Individual health risk as a function of vaccine uptake.
- Immigration propensity as a function of migrant stock.

Agents can work together to produce a joint outcome.

- Production functions.
- Wage in worker-firm problems.

Graph terminology

Let V be a set of vertices (nodes) with $|V| = n$; can normalize $V = \{1, \ldots, n\}$.

Let E be a set of edges between nodes in V. This is a set of pairs of nodes $(i, j) \in V \times V$.

Then $G = (V, E)$ is a graph.

G is (un)directed if the pairs in E or (un)ordered.

G is a simple graph if a pair (i, j) has at most one edge between them.

G is a multigraph if the latter is not the case.

Edges can carry a weight, $w_{i,j} \geq 0$, in which case G is a weighted graph.

G can be represented by its (weighted) $n \times n$ adjacency matrix $\boldsymbol{A}$.

In the unweighted case,

$$(\boldsymbol{A})_{i,j} = \begin{cases} 0 & \text{if } (i,j) \notin E \\ \\ 1 & \text{if } (i,j) \in E \end{cases}$$

In the weighted case,

$$(\boldsymbol{A})_{i,j} = \begin{cases} 0 & \text{if } (i,j) \notin E \\ \\ w_{i,j} & \text{if } (i,j) \in E \end{cases}$$

We usually exclude self links, so the diagonal of $\boldsymbol{A}$ contains only zeros.

In the undirected case, $\boldsymbol{A}$ is symmetric.

The (weighted) degree of vertex i is

$$d_i = \sum_{j=1}^{n} w_{i,j} = \sum_{j:(i,j)\in E} w_{i,j}.$$

In the unweighted case d_i is the number of (if the graph is directed, outgoing) edges involving node i.

The volume of a set of nodes $V_1 \subseteq V$ is

$$\mathrm{vol}(V_1) = \sum_{i\in V_1} d_i = \sum_{i\in V_1} \sum_{j:(i,j)\in E} w_{i,j}.$$

A subset V_1 of V is connected if any two vertices in V_1 can be joined by a path of which all the intermediate points also lie in V_1.

The subset V_1 is a connected component if V_1 is connected and there are no connections between V_1 and $\bar{V}_1 = V \backslash V_1$.

The collection of sets $V_1, \ldots, V_m$ (for some m) form a partition of G if each of them is connected and

$$V_i \cap V_j = \emptyset \text{ for all } i \neq j \qquad \text{and} \qquad V_1 \cup \ldots \cup V_m = V.$$

(For an undirected graph) the Laplacian of G is

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A},$$

for

$$\boldsymbol{D} = \text{diag}(\boldsymbol{d}) = \text{diag}(d_1, \ldots, d_n)$$

Note that, for any $\boldsymbol{v} \in \mathbb{R}^n$,

$$\begin{aligned}\boldsymbol{v}' \boldsymbol{L}\, \boldsymbol{v} &= \sum_{i=1}^{n} v_i^2\, d_i - \sum_{i=1}^{n}\sum_{j=1}^{n} v_i\, v_j\, w_{i,j} = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i^2\, w_{i,j} - \sum_{i=1}^{n}\sum_{j=1}^{n} v_i\, v_j\, w_{i,j} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i\, (v_i - v_j)\, w_{i,j} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} (v_i - v_j)^2\, w_{i,j},\end{aligned}$$

where we have used the accounting identity $(v_i - v_j)^2 = v_i(v_i - v_j) + v_j(v_j - v_i)$.

So, $\boldsymbol{L}$ is positive semi-definite. It is singular because

$$\boldsymbol{L}\, \mathbf{1}_n = \boldsymbol{D}\, \mathbf{1}_n - \boldsymbol{A}\, \mathbf{1}_n = \boldsymbol{d} - \boldsymbol{d} = 0\, \mathbf{1}_n$$

and so its smallest eigenvalue is zero (with eigenvector $\mathbf{1}_n$).

The number of connected components in graph G equals the multiplicity of the eigenvalue zero of $\boldsymbol{L}$.

Consider first a connected graph. Let $\boldsymbol{v}$ be an eigenvector of $\boldsymbol{L}$ associated with eigenvalue 0. Then

$$0 = \boldsymbol{v}' \boldsymbol{L} \boldsymbol{v} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (v_i - v_j)^2 \, w_{i,j}.$$

As $w_{i,j} \geq 0$ this sum can only be zero if $(v_i - v_j)^2 \, w_{i,j} = 0$ for all $(i,j) \in E$. For such (i,j) we necessarily have $v_i = v_j$. Further, any third node k that can be reached via the path $i \to j \to k$ then also has $v_k = v_j = v_i$. Extending the argument to longer paths and invoking that the graph is connected implies that $\boldsymbol{v} = \mathbf{1}_n$.

Next take m connected components. Then G is the collection of m connected subgraphs $G_1, \ldots G_m$, each with their own proper Laplacian, $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_m$. To each of these the above argument applies. Further, we may always re-arrange the nodes such that $\boldsymbol{L} = \text{blockdiagonal}(\boldsymbol{L}_1, \ldots, \boldsymbol{L}_m)$. The eigenvectors of $\boldsymbol{L}$ are the eigenvectors of the $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_m$ (supplemented with zeros as additional entries), so that the eigenvalue 0 has m distinct eigenvectors associated with it; these are $\mathbf{1}_{V_1}, \ldots, \mathbf{1}_{V_m}$, where $\mathbf{1}_V$ has the n entries $(\mathbf{1}_V)_i = \{i \in V\}$. □

The normalized adjacency matrix is

$$\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{A}$$

(the notation presumes that G is connected; otherwise use the Moore-Penrose pseudo inverse).

Note that

$$(\boldsymbol{P})_{i,j} = \begin{cases} 0 & \text{if } (i,j) \notin E \\ w_{i,j}/d_i & \text{if } (i,j) \in E \end{cases}$$

and that the rows of $\boldsymbol{P}$ all sum up to one.

Row i of $\boldsymbol{P}$ gives the probability distribution of taking a random step through the graph $\boldsymbol{G}$ when starting at node i.

$\boldsymbol{P}$ is the transition matrix of a Markov chain on G.

If G is connected (and not bipartite) then $\boldsymbol{P}$ has a steady-state distribution given by

$$p_i = \frac{d_i}{\text{vol}(V)}$$

Can define (at least) two normalized version of the Laplacian matrix $\boldsymbol{L}$.

A symmetric normalization:

$$\boldsymbol{L}_{\text{sym}} = \boldsymbol{D}^{-1/2}\boldsymbol{L}\,\boldsymbol{D}^{-1/2} = \boldsymbol{I}_n - \boldsymbol{D}^{-1/2}\boldsymbol{A}\,\boldsymbol{D}^{-1/2}$$

and a 'random-walk' normalization:

$$\boldsymbol{L}_{\text{rw}} = \boldsymbol{D}^{-1}\boldsymbol{L} = \boldsymbol{I}_n - \boldsymbol{P}.$$

In both cases, can show similar properties as for $\boldsymbol{L}$ before.

A connection between the two is as follows:

$\lambda \geq 0$ is an eigenvalue of $\boldsymbol{L}_{\text{rw}}$ with eigenvector $\boldsymbol{v}$ if and only if λ is an eigenvalue of $\boldsymbol{L}_{\text{sym}}$ with eigenvector $\boldsymbol{D}^{1/2}\boldsymbol{v}$.

The number of zero eigenvalues continues to give the number of connected components in G.

For a connected graph G the second smallest eigenvalue $\lambda_2 > 0$ is a measure of global connectivity of G.

This spectral gap is connected to the Cheeger constant

$$C = \min_{\{V_1 \subset V : 0 < \sum_{i \in V_1} d_i \leq \sum_{i \in \bar{V}_1} d_i\}} \frac{\sum_{i \in V_1} \sum_{j \in \bar{V}_1} w_{i,j}}{\sum_{i \in V_1} d_i}$$

through the inequalities

$$1/2 C^2 \leq \lambda_2 \leq 2C.$$

$C \in [0, 1]$ measures how difficult it is to separate G into two disconnected components by removing edges from it.

The numerator in the definition of C is the total weight of the removed edges, the denominator is the total degree in the smallest of the two components.

Small positive C show that there is a bottleneck in G, and that the graph is only weakly connected.

Outline

Motivation

Aim:

Partition set of nodes into m clusters (groups) so that they are similar within a cluster and heterogeneous across clusters.

Here, consider undirected (non-bipartite) graphs.

m is taken as given.

Several types of clustering exist. Here, a flavor is given.

We will discuss spectral clustering based on the Laplacian and on the (random-walk) normalized Laplacian.

Unnormalized clustering

Consider clustering V into clusters $V_1, \ldots, V_m$.

One motivation: minimize the ratio cut of G:

$$\text{RatioCut}(V_1, \ldots, V_m) = \sum_{k=1}^{m} \frac{\text{cut}(V_k, \bar{V}_k)}{|V_k|}$$

for

$$\text{cut}(V_k, \bar{V}_k) = \frac{1}{2} \sum_{i \in V_k} \sum_{j \in \bar{V}_k} w_{i,j} + \frac{1}{2} \sum_{i \in \bar{V}_k} \sum_{j \in V_k} w_{i,j} = \sum_{i \in V_k} \sum_{j \in \bar{V}_k} w_{i,j}$$

(using undirectedness in the last transition).

This minimization problem tries to minimize the weight assigned to edges that go across different clusters.

The weighting makes larger clusters more attractive.

For $k = 1, \ldots, m$, introduce the n-vectors

$$(\boldsymbol{h}_k)_i = \begin{cases} {}^1\!/\sqrt{|V_k|} & \text{if } i \in V_k \\ 0 & \text{if } i \notin V_k \end{cases}.$$

For each $k = 1, \ldots, m$ we have (from calculations given above) that

$$\begin{aligned} \boldsymbol{h}_k' \boldsymbol{L}\, \boldsymbol{h}_k &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} ((\boldsymbol{h}_k)_i - (\boldsymbol{h}_k)_j)^2\, w_{i,j} \\ &= \frac{1}{2} \sum_{i \in V_k} \sum_{j \notin V_k} \frac{1}{|V_k|}\, w_{i,j} + \frac{1}{2} \sum_{i \notin V_k} \sum_{j \in V_k} \frac{1}{|V_k|}\, w_{i,j} \\ &= \frac{\mathrm{cut}(V_k, \bar{V}_k)}{|V_k|}. \end{aligned}$$

Therefore,

$$\mathrm{RatioCut}(V_1, \ldots, V_m) = \sum_{k=1}^{m} \frac{\mathrm{cut}(V_k, \bar{V}_k)}{|V_k|} = \mathrm{trace}\, \boldsymbol{H}' \boldsymbol{L}\, \boldsymbol{H}$$

where $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_m)$.

Note that, for $k, \ell \in \{1, \ldots, m\}^2$

$$\boldsymbol{h}_k' \boldsymbol{h}_\ell = \sum_{i=1}^{n} (\boldsymbol{h}_k)_i \, (\boldsymbol{h}_\ell)_i = \begin{cases} 1 & \text{if } k = \ell \\ 0 & \text{if } k \neq \ell \end{cases},$$

so that $\boldsymbol{H}' \boldsymbol{H} = \boldsymbol{I}_m$ meaning that $\boldsymbol{H}$ is an $n \times m$ orthonormal matrix.

Minimizing the RatioCut can then be written as the discrete optimization problem

$$\min_{V_1, \ldots, V_m} \operatorname{trace} \boldsymbol{H}' \boldsymbol{L} \boldsymbol{H}$$

subject to $\boldsymbol{H}$ being orthonormal and being of the form given on the previous slide.

This problem is NP hard.

A relaxation of the discrete problem is to solve

$$\min_{\boldsymbol{H}} \operatorname{trace} \boldsymbol{H}' \boldsymbol{L} \boldsymbol{H}$$

subject to $\boldsymbol{H}$ being orthonormal (but letting its columns take on arbitrary values in $\mathbb{R}^n$)

This is a well-known problem.

The solution is that $\boldsymbol{H}$ are the eigenvectors associated with the m smallest eigenvalues of $\boldsymbol{L}$.

Let $\boldsymbol{q}_i$ be the ith row of $\boldsymbol{H}$. This is not a classifier.

To obtain one we solve

$$\min_{V_1,\ldots,V_m} \sum_{k=1}^{m} \frac{\sum_{i \in V_k, j \in V_k} \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2}{|V_k|} = \min_{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_m} \sum_{i=1}^{n} \min_{g} \|\boldsymbol{q}_i - \boldsymbol{\mu}_g\|^2$$

which minimizes the within-cluster sum of squares. This is the k-means algorithm.

This step is ad hoc without further justification (!)

Normalized clustering

An alternative to RatioCut is

$$\mathrm{Ncut}(V_1, \ldots, V_m) = \sum_{k=1}^{m} \frac{\mathrm{cut}(V_k, \bar{V}_k)}{\mathrm{vol}(\mathrm{V_k})},$$

which features a different normalization in the summands.

Recall that we want to look for a partitioning so that (i) edges within a cluster have a high weight and (ii) edges across different clusters have a low weight.

(ii) amounts to making $\mathrm{cut}(V_k, \bar{V}_k)$ small.

(i) amounts to making

$$\sum_{(i,j)\in V_k} w_{i,j} = \sum_{i\in V_k} \sum_{j=1}^{n} w_{i,j} - \sum_{i\in V_k} \sum_{j\in \bar{V}_k} w_{i,j} = \mathrm{vol}(V_k) - \mathrm{cut}(V_k, \bar{V}_k)$$

large. This is done by both minimizing the cut and maximizing the volume.

Ncut achieves both (i) and (ii) while Ratiocut only looks at (ii).

Like RatioCut, the Ncut problem has a connection to the Laplacian matrix.

Redefine the columns of $\boldsymbol{H}$ as

$$(\boldsymbol{h}_1)_i = \begin{cases} 1/\sqrt{\mathrm{vol}(V_k)} & \text{if } i \in V_k \\ 0 & \text{if } i \notin V_k \end{cases} .$$

Then

$$\mathrm{Ncut}(V_1, \ldots, V_m) = \mathrm{trace}\, \boldsymbol{H}' \boldsymbol{L}\, \boldsymbol{H}$$

subject to $\boldsymbol{H}' \boldsymbol{D} \boldsymbol{H} = \boldsymbol{I}_m$ and $\boldsymbol{H}$ being of the form given above.

Can substitute in $\dot{\boldsymbol{H}} = \boldsymbol{D}^{1/2} \boldsymbol{H}$ to obtain the alternative representation

$$\mathrm{Ncut}(V_1, \ldots, V_m) = \mathrm{trace}\, \dot{\boldsymbol{H}}' \boldsymbol{D}^{-1/2} \boldsymbol{L}\, \boldsymbol{D}^{-1/2} \dot{\boldsymbol{H}} = \mathrm{trace}\, \dot{\boldsymbol{H}}' \boldsymbol{L}_{\mathrm{sym}}\, \dot{\boldsymbol{H}}$$

subject to $\dot{\boldsymbol{H}}' \dot{\boldsymbol{H}} = \boldsymbol{I}_m$ and $\dot{\boldsymbol{H}}$ being of the form given above.

Hence, relaxing discreteness yields the solution that $\dot{\boldsymbol{H}}$ are the eigenvectors of $\boldsymbol{L}_{\mathrm{sym}}$ associated with the m smallest eigenvalues.

But then $\boldsymbol{H}$ are the corresponding eigenvectors of $\boldsymbol{L}_{\mathrm{rw}}$.

An alternative view on Ncut comes from the following observation.

Let $\{X_t\}$ be a stationary Markov process on G. Then

$$\mathrm{Ncut}(V_k, \bar{V}_k) = \Pr(X_t \in V_k | X_{t-1} \in \bar{V}_k) + \Pr(X_t \in \bar{V}_k | X_{t-1} \in V_k).$$

This is the probability of switching to/from a cluster.

Normalized spectral clustering aims to make the probability of such an event small.

A proof of the above comes from the observation that, for any set $A \subset V$,

$$\Pr(X_t \in A, X_{t-1} \in \bar{A}) = \sum_{i \in \bar{A}} \sum_{j \in A} \Pr(X_t = j, X_{t-1} = i) = \sum_{i \in \bar{A}} \sum_{j \in A} \pi_i \boldsymbol{P}_{i,j}$$

but

$$\pi_i = \frac{d_i}{\mathrm{vol}(V)}, \qquad \boldsymbol{P}_{i,j} = \frac{w_{i,j}}{d_i},$$

and so

$$\Pr(X_t \in A, X_{t-1} \in \bar{A}) = \frac{1}{\mathrm{vol}(V)} \sum_{i \in \bar{A}} \sum_{j \in A} w_{i,j}.$$

Therefore,

$$\Pr(X_t \in A | X_{t-1} \in \bar{A}) = \frac{\Pr(X_t \in A, X_{t-1} \in \bar{A})}{\Pr(X_{t-1} \in \bar{A})} = \frac{1}{\mathrm{vol}(\bar{A})} \sum_{i \in \bar{A}} \sum_{j \in A} w_{i,j},$$

which equals

$$\frac{\mathrm{cut}(\bar{A}, A)}{\mathrm{vol}(\bar{A})}.$$

Reversing the roles of A and $\bar{A}$ in the above and collecting terms yields the result.

Outline

Classic SBM

Consider an unweighted undirected graph G. (If weighted, can always reduce to binary case.)

Suppose that nodes are of one of m latent types.

Can think about this as (independent) latent variables α_i that take values from the set $\{1, \ldots, m\}$.

Conditional on the types of all nodes, edges are placed independently.

The probability of an edge between (i, j) depends on α_i, α_j.

Thus

$$\mathbb{P}(\boldsymbol{A}) = \sum_{(z_1,\ldots,z_n)\in\{1,\ldots,m\}^n} \mathbb{P}(\boldsymbol{A}, \alpha_1 = z_1, \ldots, \alpha_n = z_n)$$

where the summand factors as

$$\prod_{i=1}^{n} \mathbb{P}(\alpha_i = z_i) \prod_{i<j} \mathbb{P}((\boldsymbol{A})_{i,j} | \alpha_i = z_i, \alpha_j = z_j).$$

This is a simple but flexible model that allows for (dis-)assortative matching between nodes.

Parametrization of heterogeneity is low dimensional:

$m \times m$ symmetric matrix $\boldsymbol{B}$ of bernoulli success probabilities for pairs of types:

$$(\boldsymbol{B})_{z_1,z_2} = (\boldsymbol{B})_{z_2,z_1} = \Pr(w_{i,j} = 1 | \alpha_i = z_1, \alpha_j = z_2)$$

and an m-vector of type probabilities.

In this model there are m latent communities.

Spectral clustering correctly classifies nodes to communities as $n \to \infty$.

Spectral clustering in SBM (fixed effects)

The intuition behind the consistency of spectral clustering has two steps:

(i) The eigenvectors of the (normalized) Laplacian converge in probability to those of the expected Laplacian.

(ii) The expected Laplacian has only m eigenvectors, which identify group membership.

Point (i) is a technical issue. We focus on Point (ii).

We do this conditional on the variables $\alpha_1, \ldots, \alpha_n$; so we treat these as fixed.

Note that the $n \times n$ matrix $\boldsymbol{A}$ has (conditional) expectation

$$\boldsymbol{A}_0 = \mathbb{E}(\boldsymbol{A}|\alpha_1, \ldots, \alpha_n) = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}'$$

for $n \times m$ matrix $\boldsymbol{Z}$ whose ith column is composed of all zeros expect at location α_i.

Will assume that all columns of $\boldsymbol{Z}$ contain at least one 1 (so all communities have at least one member, and so the matrix has full column rank) and that $\boldsymbol{B}$ has full rank.

The normalized Laplacian associated with $\boldsymbol{A}_0$ is

$$\boldsymbol{L}_0 = \boldsymbol{I}_n - \boldsymbol{D}_0^{-1/2} \boldsymbol{A}_0 \boldsymbol{D}_0^{-1/2}$$

for $\boldsymbol{D}_0$ the diagonal degree matrix.

$\boldsymbol{L}_0$ and $\bar{\boldsymbol{L}}_0 = \boldsymbol{I}_n - \boldsymbol{L}_0$ have the same eigenvectors so focus on the latter here.

Note that $\boldsymbol{D}_0 = \text{diag}(\boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}'\mathbf{1}_n)$ and, therefore, that

$$(\boldsymbol{D}_0)_{i,i} = \sum_{z=1}^{m} \boldsymbol{B}_{Z_i,z}\, n_z$$

for $n_z = \sum_{i=1}^{n} \{\alpha_i = z\}$. Hence, the entries of the $n \times n$ diagonal matrix $\boldsymbol{D}_0$ can only take on m different values. Collect these values in the $m \times m$ diagonal matrix

$$(\bar{\boldsymbol{D}}_0)_{z,z} = \sum_{z'=1}^{m} \boldsymbol{B}_{z,z'}\, n_{z'}.$$

Then

$$\bar{\boldsymbol{L}}_0 = \boldsymbol{Z}\bar{\boldsymbol{D}}_0^{-1/2}\boldsymbol{B}\bar{\boldsymbol{D}}_0^{-1/2}\boldsymbol{Z}' = \boldsymbol{Z}\bar{\boldsymbol{B}}\boldsymbol{Z}'$$

for $\bar{\boldsymbol{B}} = \bar{\boldsymbol{D}}_0^{-1/2}\boldsymbol{B}\bar{\boldsymbol{D}}_0^{-1/2}$.

Next, consider the (small) $m \times m$ matrix

$$\boldsymbol{C}\bar{\boldsymbol{B}}\boldsymbol{C}$$

for $\boldsymbol{C} = (\boldsymbol{Z}'\boldsymbol{Z})^{1/2}$. Note that $\boldsymbol{C}\bar{\boldsymbol{B}}\boldsymbol{C}$ is a full rank and symmetric. So we have the eigendecomposition

$$\boldsymbol{C}\bar{\boldsymbol{B}}\boldsymbol{C} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}'$$

where all m eigenvalues are non-zero.

With $\boldsymbol{C}$ invertible, we have $\bar{\boldsymbol{B}} = \boldsymbol{C}^{-1}\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}'\boldsymbol{C}^{-1}$ and, therefore, also that

$$\bar{\boldsymbol{L}}_0 = \boldsymbol{Z}\bar{\boldsymbol{B}}\boldsymbol{Z}' = \boldsymbol{Z}(\boldsymbol{C}^{-1}\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}'\boldsymbol{C}^{-1})\boldsymbol{Z}' = \boldsymbol{Z}\boldsymbol{Q}\,\boldsymbol{S}\,\boldsymbol{Q}'\boldsymbol{Z}'$$

for $\boldsymbol{Q} = \boldsymbol{C}^{-1}\boldsymbol{V}$.

Note that $(\boldsymbol{Z}\boldsymbol{Q})'(\boldsymbol{Z}\boldsymbol{Q}) = \boldsymbol{Q}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{Q} = \boldsymbol{V}'(\boldsymbol{Z}'\boldsymbol{Z})^{-1/2}(\boldsymbol{Z}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1/2}\boldsymbol{V} = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_m$. So the columns of the $n \times m$ orthonormal matrix $\boldsymbol{Z}\boldsymbol{Q}$ are the eigenvectors of the matrix $\bar{\boldsymbol{L}}_0$ (and so equally of $\boldsymbol{L}_0$).

Finally, letting $\boldsymbol{z}_i$ the ith row of $\boldsymbol{Z}$, we have that $\boldsymbol{z}_i\,\boldsymbol{Q} = \boldsymbol{z}_j\,\boldsymbol{Q}$ if and only if

$$\boldsymbol{z}_i = \boldsymbol{z}_j \qquad (\text{and so } \alpha_i = \alpha_j)$$

because $\boldsymbol{Q}$ is invertible. Hence, clustering on the eigenvectors of $\boldsymbol{L}_0$ yields correct assignment of nodes to communities.

Consistency of spectral clustering has been established in more general stochastic block models.

This includes the setting where m grows slowly with n, and also several forms of sparsity.

There is not a single definition of a sparse graph.

Roughly, sparse means that the number of edges does not grow proportional to n^2 (which would give rise to a dense network). This is often formulated in the (expected) degree d_i growing only slowly with n.

In the stochastic block model, usually done by including a (common) sparsity parameter that shrinks the probability of edge formation to zero as n grows.

A weighted graph can always be translated into an unweighted graph so the above equally applies to the weighted case.

In principle, with a consistent classification in hand, we can treat cluster membership as known and then estimate the model parameters by standard methods.

Identification in small networks (random effects)

We can see the stochastic block model as a (nonlinear) model with two-way heterogeneity:

$$(\boldsymbol{A})_{i,j} = w_{i,j} = \varphi(\alpha_i, \alpha_j, \varepsilon_{i,j}),$$

where the α_i are i.i.d. across i and the $\varepsilon_{i,j}$ are i.i.d. across dyads, independent of the α_i.

Similar to a finite-mixture model in panel data.

Edge weights are dependent through their joint dependence on the α_i only.

By using such conditional-independence we can nonparametrically identify

The distribution of $w_{i,j}|\alpha_i, \alpha_j$ and,

the distribution of α_i,

from the joint distribution of edge weights in small graphs.

Easiest to see under the assumption that the functions

$$G_z(w) = \mathbb{P}(w_{i,j} \leq w | \alpha_i = z) = \sum_{z'=1}^{m} \mathbb{P}(w_{i,j} \leq w | \alpha_i = z, \alpha_j = z')\, p_{z'}$$

are linearly independent.

In that case, $n \geq 4$ suffices for identification.

Consider a grid of m points $w_1, \ldots w_m$ and construct the $m \times m$ matrix

$$\boldsymbol{G} = \begin{pmatrix} G_1(w_1) & \ldots & G_m(w_1) \\ \vdots & \ldots & \vdots \\ G_1(w_m) & \ldots & G_m(w_m) \end{pmatrix}.$$

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

Next, for distinct integers $i, j, k, \ell \in \{1, \ldots, n\}^4$, let

$$(\boldsymbol{M}\ \ )_{m_1,m_2} = \mathbb{P}(w_{i,j} \leq w_{m_1}, w_{i,k} \leq w_{m_2}),$$
$$(\boldsymbol{M}_w)_{m_1,m_2} = \mathbb{P}(w_{i,j} \leq w_{m_1}, w_{i,k} \leq w_{m_2}, w_{i,\ell} \leq w),$$

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

Then conditional independence implies that

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

and

$$\boldsymbol{M}_w = \boldsymbol{G}\boldsymbol{D}^{1/2}\boldsymbol{D}_w\boldsymbol{D}^{1/2}\boldsymbol{G}'$$

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

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

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

Note that

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

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

Next,

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

for any w.

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 $\boldsymbol{V}\boldsymbol{M}_{w_1}\boldsymbol{V}^\top, \ldots, \boldsymbol{V}\boldsymbol{M}_{w_m}\boldsymbol{V}^\top$.

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.

Next, construct the matrix

$$(\boldsymbol{N}_w)_{m_1,m_2} = \mathbb{P}(w_{i,k} \leq w_{m_1}, w_{j,\ell} \leq w_{m_2}, w_{i,j} \leq w).$$

Then

$$\boldsymbol{N}_w = \boldsymbol{G}\boldsymbol{D}\boldsymbol{H}_w\boldsymbol{D}\boldsymbol{G}'$$

for

$$(\boldsymbol{H}_w)_{z,z'} = \mathbb{P}(w_{i,j} \leq w | \alpha_i = z, \alpha_j = z').$$

But

$$\boldsymbol{V}\boldsymbol{N}_w\boldsymbol{V}^\top = (\boldsymbol{V}\boldsymbol{G}\boldsymbol{D}^{1/2})\boldsymbol{D}^{1/2}\boldsymbol{H}_w\boldsymbol{D}^{1/2}(\boldsymbol{D}^{1/2}\boldsymbol{G}'\boldsymbol{V}^\top) = \boldsymbol{Q}\boldsymbol{D}^{1/2}\boldsymbol{H}_w\boldsymbol{D}^{1/2}\boldsymbol{Q}'$$

and so the entries of

$$\bar{\boldsymbol{Q}}'\boldsymbol{V}\boldsymbol{N}_w\boldsymbol{V}^\top\bar{\boldsymbol{Q}} = \boldsymbol{\Sigma}'\boldsymbol{\Delta}(\boldsymbol{D}^{1/2}\boldsymbol{H}_w\,\boldsymbol{D}^{1/2})\boldsymbol{\Delta}\boldsymbol{\Sigma}$$

give the

$$\mathbb{P}(w_{i,j} \leq w | \alpha_i = z, \alpha_j = z')\,\sqrt{p_z\,p_{z'}}$$

up to sign and labelling. Note that for $w = +\infty$ this gives $\sqrt{p_z\,p_{z'}}$ up to the same sign and labelling, so that the ratio yields $\mathbb{P}(w_{i,j} \leq w | \alpha_i = z, \alpha_j = z')$ while the diagonals yield the p_z.

Outline

Models on networks

If, in an (undirected) network of n agents we observe interactions between all

$$\frac{n(n-1)}{2}$$

possible pairs of nodes we can consider estimation of nonlinear fixed-effect models.

Say

$$y_{i,j}|x_{i,j}, \alpha_i, \alpha_j$$

has a parametric distribution. Can do bias-corrected maximum likelihood estimation.

This is like a panel data problem where $N, T \to \infty$ at the same rate; for each of n nodes we observe $(n-1)$ outcomes, so we get more measurements on each individual as $n \to \infty$.

(Fixed-effect estimation from many small networks will not be possible.)

The case where we observe more limited interaction between nodes is much more complicated.

A linear model

Consider an undirected (multi-) graph where each edge (i, j) has associated with it an outcome $y_{i,j}$, with

$$\mathbb{E}(y_{i,j}|\alpha_i, \alpha_j, w_{i,j} > 0) = \alpha_i - \alpha_j$$

Say we observe m edges.

Stack all outcomes in m-vector $\boldsymbol{y}$.

Let $\boldsymbol{B}$ be the $m \times n$ matrix where the row for edge $e = (i, j)$ has $(\boldsymbol{B})_{e,i} = 1$ and $(\boldsymbol{B})_{e,j} = -1$. This is the (oriented) incidence matrix of the graph; we note that

$$\boldsymbol{L} = \boldsymbol{B}'\boldsymbol{B}$$

connects it to the Laplacian matrix.

We can write

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \qquad \mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{B}) = \boldsymbol{0}.$$

The conditional-mean restriction is one of network exogeneity.

A bipartite special case

Suppose that the node set V can be partitioned into two sets V_1, V_2 and that edges are only formed between the two subsets.

Can then reparametrize

$$\alpha_i = \begin{cases} \eta_i & \text{if } i \in V_1 \\ -\gamma_i & \text{if } i \in V_2 \end{cases}$$

An example is worker i works in firm j at time t; so that log-wage $y_{i,j,t}$ decomposes as

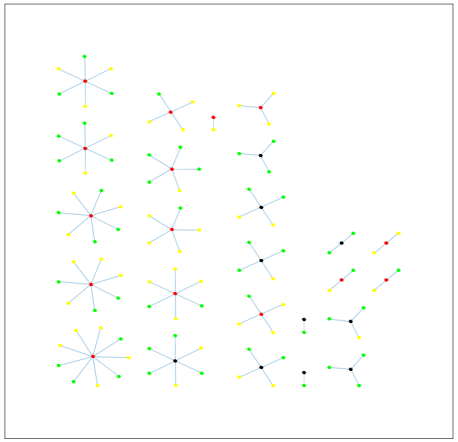
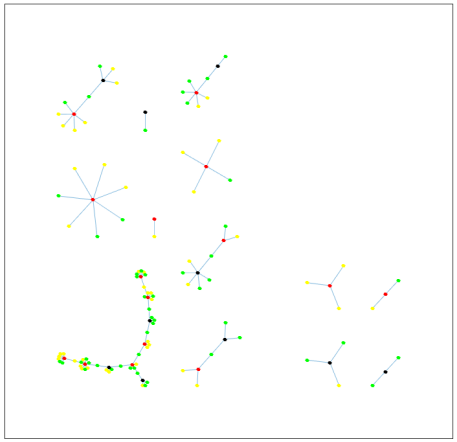
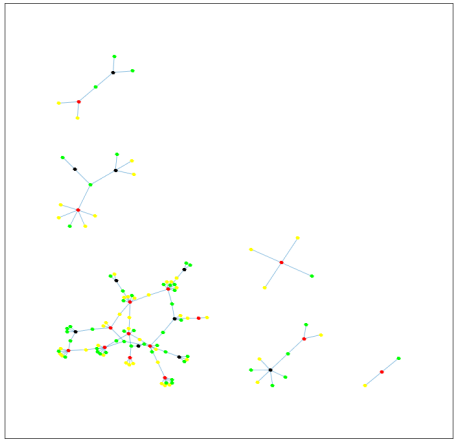
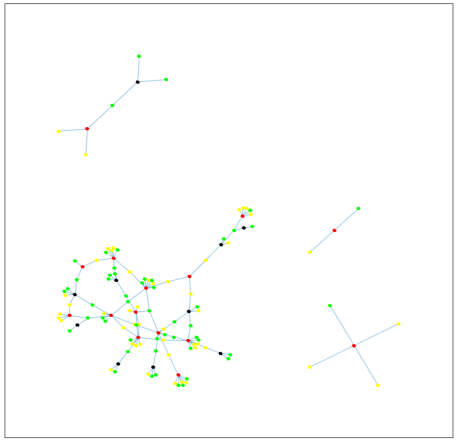
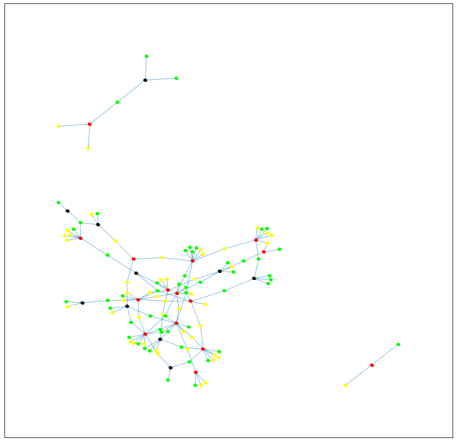
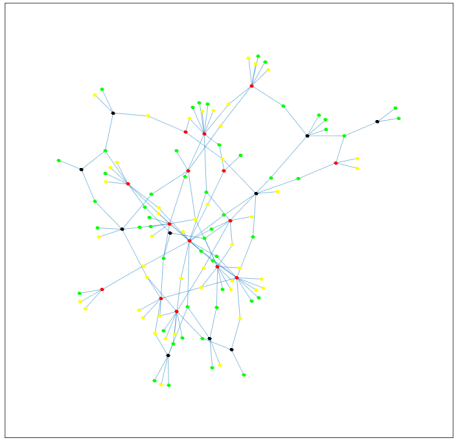
$$y_{i,j,t} = \eta_i + \gamma_{j(i,t)} + \varepsilon_{i,j,t}.$$

Here, the number of edges between any i and j is the number of periods that i is employed by j.

The connectivity structure of the induced graph drives statistical properties of the fixed-effect estimator.

For example, if no worker ever switches employer we cannot learn anything.

(Dis-)connected graphs

The above plot is for a stationary model where workers and firms are of binary types (high and low quality), and mobility of workers is introduced through exogenous lay-offs.

A one-period network has as many components as firms.

Cannot disentangle worker effects from firm effects.

Adding time periods makes workers switch to different firms.

Makes the graph more connected.

With enough mobility, the graph becomes connected.

Connectivity allows for the least-squares estimator to be well defined (subject to one normalization on the fixed effects) but is not enough for it to have good statistical properties.

Least-squares estimator

Each row of $\boldsymbol{B}$ sums to zero so we normalize the fixed effects so that

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(\boldsymbol{A})_{i,j}\,(\alpha_i+\alpha_j)=0.$$

If the graph is connected then the (constrained) least-squares estimator exists an is equal to

$$\hat{\boldsymbol{\alpha}}=\boldsymbol{D}^{-1/2}(\boldsymbol{L}_{\text{sym}})^{*}\,\boldsymbol{D}^{-1/2}\,\boldsymbol{B}'\boldsymbol{y}.$$

By strict exogeneity this estimator is unbiased.

Further note that

$$\text{var}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}|\boldsymbol{B})=\boldsymbol{D}^{-1/2}(\boldsymbol{L}_{\text{sym}})^{*}\,\boldsymbol{D}^{-1/2}\,\boldsymbol{B}'\,\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{B})\,\boldsymbol{B}\boldsymbol{D}^{-1/2}(\boldsymbol{L}_{\text{sym}})^{*}\,\boldsymbol{D}^{-1/2}.$$

Notably, with $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \, \boldsymbol{I}_m)$,

$$\mathrm{var}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} | \boldsymbol{B}) = \sigma^2 \, \boldsymbol{D}^{-1/2} (\boldsymbol{L}_{\mathrm{sym}})^* \, \boldsymbol{D}^{-1/2}.$$

So the precision of the estimator is driven by the Laplacian of the graph.

We have

$$\mathrm{var}(\hat{\alpha}_i | \boldsymbol{B}) = \sigma^2 \, \frac{(\boldsymbol{L}^*_{\mathrm{sym}})_{i,i}}{d_i}$$

The behavior of this variance depends in a very complicated way on the connectivity structure of the network.

No reason it would behave proportional to d_i^{-1}, which would correspond to the usual parametric rate.

Let λ_2 be second smallest eigenvalue of $\boldsymbol{L}_{\text{sym}}$ and

$$h_i = \left(\frac{1}{d_i} \sum_{j=1}^{n} \frac{w_{i,j}^2}{d_j} \right)^{-1}.$$

Can derive that

$$\frac{\sigma^2}{d_i} - \frac{2\,\sigma^2}{m} \leq \text{var}(\hat{\alpha}_i | \boldsymbol{B}) \leq \frac{\sigma^2}{d_i} \left(1 + \frac{1}{\lambda_2 h_i} \right) - \frac{2\,\sigma^2}{m}.$$

So,

$$\text{var}(\hat{\alpha}_i) = \frac{\sigma^2}{d_i} + o(d_i^{-1})$$

provided that $\lambda_2 h_i \to \infty$.

In this case can get conventional asymptotic normality with d_i serving as effective sample size for estimating α_i.

Example: student/teacher data

Pupils in Grades 4 and 5 of elementary school over the period 2008–2012.

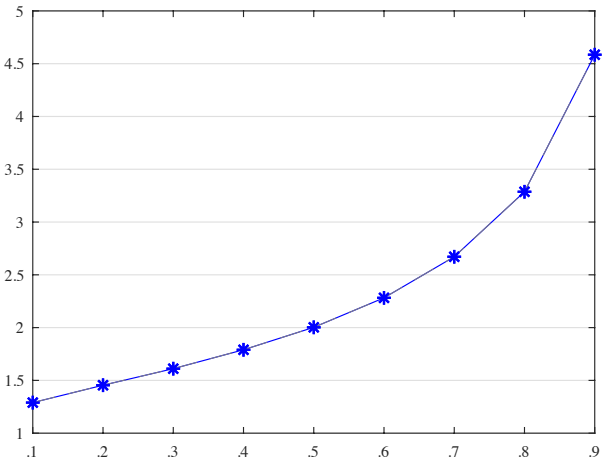
(One-mode projection has) $m = 41,612$ edges and $n = 11,945$ teachers.

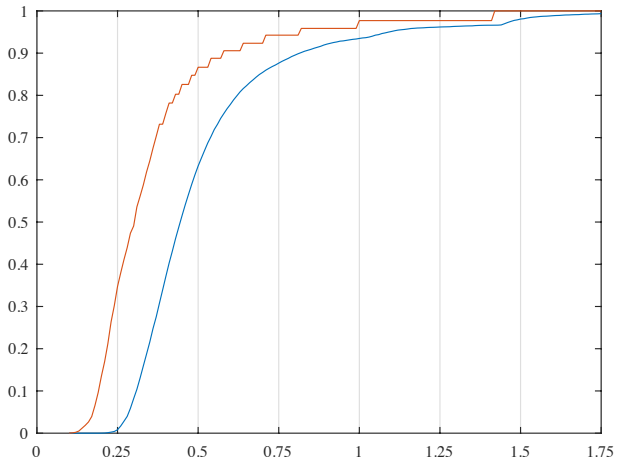
$\lambda_2 = .0039$.

	mean	stdev	$10^{\text{th}}\%$	$20^{\text{th}}\%$	$30^{\text{th}}\%$	$40^{\text{th}}\%$	$50^{\text{th}}\%$	$60^{\text{th}}\%$	$70^{\text{th}}\%$	$80^{\text{th}}\%$	$90^{\text{th}}\%$
d_i	13.87	10.76	3.00	5.50	7.50	9.00	11.00	14.00	17.50	21.50	27.50
h_i	7.15	7.13	2.43	3.30	4.01	4.72	5.48	6.36	7.44	9.12	12.56

Conventional inference procedures not applicable.

Under homokedasticity we can compute the ratio of the exact variance versus its first-order approximation.





Left plot: deciles of the distribution of this ratio.

Right plot: distribution of width of 95% confidence intervals constructed using exact (blue) and approximate (red) standard error (using $\sigma^2 = 1$).

Variance components

Interest often lies in variances and covariances between fixed effects.

These are of the form

$$\boldsymbol{\alpha}' \boldsymbol{W} \boldsymbol{\alpha}$$

for some weight matrix $\boldsymbol{W}$.

Plug-in estimator is

$$\hat{\boldsymbol{\alpha}}' \boldsymbol{W} \hat{\boldsymbol{\alpha}}$$

and

$$\mathbb{E}(\hat{\boldsymbol{\alpha}}' \boldsymbol{W} \hat{\boldsymbol{\alpha}} | \boldsymbol{B}, \boldsymbol{W}) = \boldsymbol{\alpha}' \boldsymbol{W} \boldsymbol{\alpha} + \operatorname{trace}(\boldsymbol{W} \operatorname{var}(\hat{\boldsymbol{\alpha}} | \boldsymbol{B})).$$

The bias again depends on graph structure. Here, can be corrected for exactly by subtracting an unbiased estimator of the bias.

For example, with $\boldsymbol{\varepsilon} \sim (\boldsymbol{0}, \sigma^2 \boldsymbol{I}_m)$,

$$\operatorname{trace}(\boldsymbol{W} \operatorname{var}(\hat{\boldsymbol{\alpha}} | \boldsymbol{B}) = \sigma^2 \operatorname{trace}(\boldsymbol{W} \boldsymbol{D}^{-1/2} (\boldsymbol{L}_{\text{sym}})^* \boldsymbol{D}^{-1/2}),$$

and an unbiased estimator of σ^2 can easily be constructed.

Nevertheless, only in the dense-network case is the bias-corrected estimator asymptotically normal! Inference is largely an open question.

Outline

Introduction

Interest in matched worker-firm data or student-teacher data.

Typical application estimates worker ϕ_i and firm ψ_f heterogeneity in wages

$$w_{it} = \phi_i + \psi_{f(i,t)} + \varepsilon_{it}$$

as fixed effects and then uses those to estimate functionals of their joint distribution (variance, correlation).

From above, this is plagued with econometrics problems.

Specification does not align well with standard economic models which usually feature complementarity and other nonlinear effects.

Recent literature considers nonlinear models with discrete heterogeneity.

Identification analysis takes firm heterogeneity as known.

Fundamental is to be able to (consistently) assign firms to heterogeneity types.

Requires asymptotics where minimum firm size grows without bound.

Data contain many small firms. A setting where the number of workers and number of firms grow at the same rate is more realistic. But then consistent clustering is not possible.

Data structure

Panel data on workers and firms observed over time.

For each worker i and each time period t we observe

-the worker's wage, w_{it},

-an indicator, x_{it}, concerning job mobility between periods t and $t+1$.

$x_{it} = e$ if the worker stays with the same firm,

$x_{it} = s$ if employment terminates and worker matches with a new firm,

$x_{it} = c$ if the worker changes employer through job-to-job transition.

-the identity of the firm where worker i was employed at time t, say $f(i,t)$.

Specification: Initial allocation

Joint distribution of heterogeneity is

$$p(\phi, \psi) = \mathbb{P}(\phi_i = \phi, \psi_{it} = \psi)$$

for ψ_{it} shorthand notation for $\psi_{f(i,t)}$.

Workers independently draw their type with probability

$$p(\phi) = \mathbb{P}(\phi_i = \phi) = \int p(\phi, \psi)\, d\psi.$$

Firms, in turn, draw their type independently according to

$$p(\psi) = \mathbb{P}(\psi_f = \psi) = \int p(\phi, \psi)\, d\phi.$$

Then assign a worker of type ϕ to a firm of type ψ with probability

$$p_\phi(\psi) = \frac{p(\phi, \psi)}{p(\phi)}$$

to obtain initial allocation

Specification: Wages and mobility decisions

First period wages w_{i1} are drawn from the conditional distribution $Q_{\phi_i,\psi_{i1}}$, where

$$Q_{\phi,\psi}(w) = \mathbb{P}(w_{it} \leq w | \phi_i = \phi, \psi_{it} = \psi),$$

independently for each worker.

Next, the match quality between the worker and his current firm is evaluated.

Worker draws x_{i1} from $r_{\phi_i,\psi_{i1}}$, where

$$r_{\phi,\psi}(x) = \mathbb{P}(x_{it} = x | \phi_i = \phi, \psi_{it} = \psi), \qquad x \in \{s, c, e\}.$$

There are then three possible paths into the next period.

Specification: Evolution

$\underline{x_{i1} = e}$

Then the worker remains in same firm, so $f(i,2) = f(i,1)$ and also $\psi_{i2} = \psi_{i1}$.

Within job spells wages are allowed to be Markovian, with transition kernel

$$Q_{\phi,\psi,w}(w') = \mathbb{P}(w_{it} \leq w' | w_{it-1} = w, x_{it-1} = e, \phi_i = \phi, \psi_{it} = \psi),$$

whose steady-state distribution is $Q_{\phi,\psi}$.

Thus, when $x_{i1} = e$, the second-period wage w_{i2} is a draw from $Q_{\phi_i,\psi_{i2},w_{i1}}$.

$\underline{x_{i1} = s}$

Then employment at firm $f(i, 1)$ is terminated.

Worker passes through unemployment and draws a new firm type ψ_{i2} from the conditional distribution p_{ϕ_i}.

New wage is drawn from the implied $Q_{\phi_i, \psi_{i2}}$.

Creates a 'reset' that is powerful to exploit for identification.

$\underline{x_{i1} = c}$

Worker changes employer as a consequence of on-the-job search, without passing through unemployment.

In this case, the new firm type is a draw from $k_{\phi_i,\psi_{i1}}$, for transition kernel

$$k_{\phi,\psi}(\psi') = \mathbb{P}(\psi_{it} = \psi' | \phi_i = \phi, \psi_{it-1} = \psi),$$

whose steady-state distribution is p_ϕ.

Here, we allow transitions to depend on the current employer.

This is important and introduces complicated dependence.

As in the previous case, again, the new wage is drawn from the implied $Q_{\phi_i,\psi_{i2}}$.

Assumptions

We need three assumptions.

Write

$$P_\phi(w) = \mathbb{P}(w_{it} \leq w, x_{it} = s | \phi_i = \phi) = \int r_{\phi,\psi}(s)\, Q_{\phi,\psi}(w)\, p_\phi(\psi)\, d\psi,$$

and

$$Q_\phi(w) = \int Q_{\phi,\psi}(w)\, p_\phi(\psi)\, d\psi, \qquad \text{and} \qquad Q_\psi(w) = \int Q_{\phi,\psi}(w)\, p_\psi(\phi)\, d\phi.$$

The distributions P_ϕ and Q_ϕ, and Q_ψ are linearly independent in ϕ and ψ, respectively.

For all (ϕ, ψ) it holds that

(i) $0 < p(\phi, \psi) < 1$, and

(ii) $r_{\phi,\psi}(s) > 0$.

The distributions $Q_{\phi,\psi}$ are linearly independent in ψ for each ϕ.

Identification result

Under these three assumptions:

(i) the wage distributions $Q_{\phi,\psi}$ and transition kernels $Q_{\phi,\psi,w}$,

(ii) the mobility distributions $r_{\phi,\psi}$ and transition kernels $k_{\phi,\psi}$, and

(iii) the type distribution $p(\phi, \psi)$

are all nonparametrically identified up to a relabelling of (ϕ, ψ) from a panel on wage trajectories and mobility patterns spanning three time periods.

Identification argument proceeds in multiple steps.

It exploits certain conditional-independence restrictions embedded in the model.

Identification argument: Step 1/6

First step identifies the distribution of wages conditional on the worker type alone,

$$Q_\phi(w) = \int Q_{\phi,\psi}(w)\, p_\phi(\psi)\, d\psi,$$

as well as $p(\phi)$, up to an arbitrary ordering of the ϕ.

This can be done by noting that

$$\mathbb{P}(w_{i1} \leq w_1, x_{i1} = s, w_{i2} \leq w_2, x_{i2} = s, w_{i3} \leq w_3)$$

factors as the tri-variate mixture

$$\int P_\phi(w_1)\, P_\phi(w_2)\, Q_\phi(w_3)\, p(\phi)\, d\phi,$$

where, recall, $P_\phi(w)\mathbb{P}(w_{it} \leq w, x_{it} = s|\phi_i = \phi)$, and the components are linearly independent.

Identification argument: Step 2/6

Second step identifies the distribution of wages conditional on the firm type alone, that is,

$$Q_\psi(w) = \int Q_{\phi,\psi}(w)\, p_\psi(\phi)\, d\phi,$$

as well as $p(\psi)$, again up to an arbitrary ordering of the ψ.

This, in turn, can be done by noting that

$$\mathbb{P}(w_{i_1 ft} \leq w_1, w_{i_2 ft} \leq w_2, w_{i_3 ft} \leq w_3)$$

factors as the mixture

$$\int Q_\psi(w_1)\, Q_\psi(w_2)\, Q_\psi(w_3)\, p(\psi)\, d\psi,$$

in which the components are again linearly independent.

Identification argument: Step 3/6

Third step recovers the conditional wage distributions $Q_{\phi,\psi}$ for the labelling of worker and firm types from the previous two steps.

This can be done by looking at

$$\mathbb{P}(w_{i_1 f 1} \leq w_1, w_{i_2 f 1} \leq w, x_{i_2 1} = 1, w_{i_2 2} \leq w_2),$$

which factors as

$$\iint Q_\phi(w_2)\, H(w, \phi, \psi)\, Q_\psi(w_1)\, d\phi\, d\psi,$$

where

$$H(w, \phi, \psi) = \mathbb{P}(w_{it} \leq w, x_{it} = s, \phi_i = \phi, \psi_{it} = \psi) = Q_{\phi,\psi}(w)\, r_{\phi,\psi}(s)\, p(\phi, \psi).$$

By linear independence, given knowledge of Q_ϕ and Q_ψ, we can invert this problem and solve for H.

Let

$$h(\phi,\psi) = H(+\infty,\phi,\psi) = \mathbb{P}(x_{it} = s, \phi_i = \phi, \psi_{it} = \psi) = r_{\phi,\psi}(s)\, p(\phi,\psi).$$

Then

$$Q_{\phi,\psi}(w) = \frac{H(w,\phi,\psi)}{h(\phi,\psi)}$$

is nonparametrically identified for each (ϕ,ψ) and all w, and this up to the same labelling of worker and firm types as before.

Fourth step first recovers the dynamics of wages within spells by identifying

$$Q_{\phi,\psi}(w, w') = \mathbb{P}(w_{it} \leq w, w_{it+1} \leq w' | x_{it} = e, \phi_i = \phi, \psi_{it} = \psi)$$

up to the same labelling ambiguity.

To do so look at

$$\mathbb{P}(w_{i_1 f1} \leq w_1, w_{i_2 f1} \leq w, x_{i_2 1} = e, w_{i_2 2} \leq w', x_{i_2 2} = s, w_{i_2 3} \leq w_2),$$

which factors as

$$\int\int Q_\phi(w_2)\, G(w, w', \phi, \psi)\, Q_\psi(w_1)\, d\phi\, d\psi$$

where

$$G(w, w', \phi, \psi) = Q_{\phi,\psi}(w, w')\, r_{\phi,\psi}(s)\, r_{\phi,\psi}(e)\, p(\phi, \psi).$$

By same argument

$$Q_{\phi,\psi}(w, w') = \frac{G(w, w', \phi, \psi)}{g(\phi, \psi)}$$

for $g(\phi, \psi) = G(+\infty, +\infty, \phi, \psi)$.

Identification argument: Step 5/6

In the fifth step we recover the conditional distribution of mobility decisions, $r_{\phi,\psi}$, and the joint distribution $p(\phi,\psi)$ of worker and firm types.

First, from above,

$$r_{\phi,\psi}(e) = \frac{g(\phi,\psi)}{h(\phi,\psi)}.$$

Then, given knowledge of $p(\phi)$,

$$\frac{h(\phi,\psi)}{p(\phi)} = r_{\phi,\psi}(s)\, p_\phi(\psi).$$

From this we can learn $r_{\phi,\psi}(s)$ and

$$r_{\phi,\psi}(c) = 1 - r_{\phi,\psi}(e) - r_{\phi,\psi}(s)$$

given knowledge of p_ϕ; the latter follows from a least-squares projection of

$$Q_\phi(w) = \int Q_{\phi,\psi}(w)\, p_\phi(\psi)\, d\psi$$

on $Q_{\phi,\psi}(w)$ for each ϕ.

Clearly, $p(\phi,\psi) = p_\phi(\psi)\, p(\phi)$ is then also identified.

Identification argument: Step 6/6

Only remains to identify the transition kernels $k_{\phi,\psi}$.

Let

$$M_\phi(w_1, w_2) = \mathbb{P}(w_{i1} \leq w_1, x_{i1} = c, w_{i2} \leq w_2, x_{i2} = s | \phi_i = \phi).$$

Then

$$M_\phi(w_1, w_2) = \int\int r_{\phi,\psi'}(s)\, Q_{\phi,\psi'}(w_2)\, k_{\phi,\psi}(\psi')\, r_{\phi,\psi}(c)\, Q_{\phi,\psi}(w_1)\, p_\phi(\psi)\, d\psi'\, d\psi,$$

which can be solved for $k_{\phi,\psi}$.

The M_ϕ are identified from

$$\mathbb{P}(w_{i1} \leq w_1, x_{i1} = c, w_{i2} \leq w_2, x_{i2} = s, w_{i3} \leq w_3)$$

factoring as

$$\int M_\phi(w_1, w_2)\, Q_\phi(w_3)\, p(\phi)\, d\phi.$$

Outline

Example: Vaccine efficacy

Let

$$y_i = \begin{cases} 1 & \text{if } i \text{ is infected} \\ 0 & \text{if not} \end{cases}, \qquad x_i = \begin{cases} 1 & \text{if } i \text{ is vaccinated} \\ 0 & \text{if not} \end{cases}.$$

The regression

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

ignores any effect of peers being vaccinated or being infected.

A sensible way forward is to include the vaccination propensity among peers:

$$y_i = \alpha + \beta x_i + \gamma \tilde{x}_i + \varepsilon_i, \qquad \tilde{x}_i = \sum_{j=1}^{n} \left(\frac{(\boldsymbol{A})_{i,j}}{\sum_{j'=1}^{n} (\boldsymbol{A})_{i,j'}} \right) x_j.$$

Here β is the direct effect of the vaccine, and γ captures the indirect spillover effect.

In a treatment-effect context the spillover effect is an example of interference.

If vaccination decisions are correlated among peers the simple regression will not yield the direct effect (unless $\gamma = 0$).

The extended specification still ignores the infection rate among peers.

We further generalize the specification to

$$y_i = \alpha + \delta \tilde{y}_i + \beta x_i + \gamma \tilde{x}_i + \varepsilon_i, \qquad \tilde{y}_i := \sum_{j=1}^{n} \left(\frac{(\boldsymbol{A})_{i,j}}{\sum_{j'=1}^{n} (\boldsymbol{A})_{i,j'}} \right) y_j .$$

This is an example of the so-called linear-in-means model.

A simple special case has

$$(\boldsymbol{A})_{i,j} = 1 \text{ for all } j \neq i,$$

so that everyone interacts will everyone else, but the above allows general structures of reference groups.

Peer effects

Observe data (y_i, x_i) for agents $i = 1, \ldots, n$.

Agents engage in social interactions.

Action and/or outcome of agent i may be influenced by action/outcome or a characteristic of agent j.

Corresponds to a graph G with adjacency matrix $\boldsymbol{A}$ and normalized $\boldsymbol{P}$.

Gives rise to peer effects:

Correlated effects:
Agents behave similarly because they are in the same environment.

Exogenous effects, contextual effects, or spillover effects:
The propensity of an agent to behave in a certain way varies with the distribution of characteristics in his/her reference group.

Endogenous effects:
The propensity of an agent to behave in a certain way varies with the prevalence of that behavior in the reference group.

A linear model

Stack observations in n-vectors $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{x} = (x_1, \ldots, x_n)'$.

A linear model for social interactions is

$$\boldsymbol{y} = \alpha \mathbf{1}_n + \rho \boldsymbol{P}\boldsymbol{y} + \beta \boldsymbol{x} + \delta \boldsymbol{P}\boldsymbol{x} + \boldsymbol{\varepsilon}$$

with

$$\mathbb{E}(\boldsymbol{\varepsilon} | \boldsymbol{A}, \boldsymbol{x}) = \mathbf{0}.$$

Assume that $|\rho| < 1$ so that the model has a reduced form:

$$\begin{aligned}
\boldsymbol{y} &= (\boldsymbol{I}_n - \rho \boldsymbol{P})^{-1} (\alpha \mathbf{1}_n + \beta \boldsymbol{x} + \delta \boldsymbol{P}\boldsymbol{x} + \boldsymbol{\varepsilon}) \\
&= \left(\sum_{k=0}^{\infty} \rho^k \boldsymbol{P}^k \right) (\alpha \mathbf{1}_n + \beta \boldsymbol{x} + \delta \boldsymbol{P}\boldsymbol{x} + \boldsymbol{\varepsilon}) \\
&= \frac{\alpha}{1-\rho} \mathbf{1}_n + \beta \boldsymbol{x} + \lambda \sum_{k=1}^{\infty} \rho^{k-1} \boldsymbol{P}^k \boldsymbol{x} + \sum_{k=0}^{\infty} \rho^k \boldsymbol{P}^k \boldsymbol{\varepsilon}
\end{aligned}$$

for $\lambda = \rho\beta + \delta$.

Multiplying through by $\boldsymbol{P}$ and taking conditional expectations yields

$$\mathbb{E}(\boldsymbol{P}\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{A}) = \frac{\alpha}{1-\rho}\mathbf{1}_n + \beta\boldsymbol{P}\boldsymbol{x} + \lambda(\boldsymbol{P}^2\boldsymbol{x} + \rho\boldsymbol{P}^3\boldsymbol{x} + \rho^2\boldsymbol{P}^4\boldsymbol{x} + \cdots).$$

This equation reveals that $\boldsymbol{P}^2\boldsymbol{x}$ is a relevant (and valid) instrumental variable for $\boldsymbol{P}\boldsymbol{y}$ provided that

$\lambda = \rho\beta + \delta \neq 0$ and

$\boldsymbol{I}_n, \boldsymbol{P}, \boldsymbol{P}^2$ are not colinear.

When $\rho \neq 0$ we have overidentifying restrictions coming from $\boldsymbol{P}^3\boldsymbol{x}, \boldsymbol{P}^4\boldsymbol{x}, \ldots$

Linear-in-means model

Recalling that $\boldsymbol{P}$ is a transition matrix on the network, $\boldsymbol{P}^2\boldsymbol{x}$ is an average of characteristics of peers of peers.

No-multicolinearity then requires variation in peer-group composition.

In the complete network,

$$\boldsymbol{A} = \mathbf{1}_n\mathbf{1}_n' - \boldsymbol{I}_n, \qquad \boldsymbol{D} = (n-1)\,\boldsymbol{I}_n$$

and so

$$\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{A} = \frac{1}{n-1}(\mathbf{1}_n\mathbf{1}_n' - \boldsymbol{I}_n).$$

But also

$$\boldsymbol{P}^2 = \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{A} = \frac{(n-2)}{(n-1)^2}\boldsymbol{P} + \frac{1}{(n-1)}\boldsymbol{I}_n,$$

yielding colinearity.

Variation in peer-group size

Note that

$$\sum_{j=1}^{n}(\boldsymbol{P})_{i,j}\,x_j = \frac{1}{n-1}\sum_{j\neq i}x_j = \frac{1}{n-1}(n\,\bar{x} - x_i),$$

and

$$\sum_{j=1}^{n}(\boldsymbol{P})_{i,j}\,y_j = \frac{1}{n-1}\sum_{j\neq i}y_j = \frac{1}{n-1}(n\,\bar{y} - y_i),$$

So

$$y_i = \alpha + \rho\sum_{j=1}^{n}(\boldsymbol{P})_{i,j}\,y_j + \beta x_i + \delta\sum_{j=1}^{n}(\boldsymbol{P})_{i,j}\,x_j + \frac{1}{n-1}\sum_{j\neq i}x_j + \varepsilon_i$$

is equal to

$$y_i = \alpha + \rho\frac{1}{n-1}(n\,\bar{y} - y_i) + \beta x_i + \delta\frac{1}{n-1}(n\,\bar{x} - x_i) + \frac{1}{n-1}\sum_{j\neq i}x_j + \varepsilon_i.$$

This can be re-arranged to get

$$\left(1+\rho\frac{1}{n-1}\right)y_i = \alpha + \rho\frac{n}{n-1}\bar{y} + \left(\beta+\delta\frac{1}{n-1}\right)x_i + \delta\frac{n}{n-1}\bar{x} + \varepsilon_i.$$

Averaging yields

$$\left(1+\rho\frac{1}{n-1}\right)\bar{y} = \alpha + \rho\frac{n}{n-1}\bar{y} + \left(\beta+\delta\frac{1}{n-1}\right)\bar{x} + \delta\frac{n}{n-1}\bar{x} + \bar{\varepsilon}.$$

Differencing gives

$$\left(1+\rho\frac{1}{n-1}\right)(y_i-\bar{y}) = \alpha + \left(\beta+\delta\frac{1}{n-1}\right)(x_i-\bar{x}) + (\varepsilon_i-\bar{\varepsilon}).$$

The coefficients in the (population) regression

$$(y_i-\bar{y}) = a_n + b_n(x_i-\bar{x}) + (\varepsilon_i-\bar{\varepsilon})$$

are identified an equal

$$a_n = \frac{\alpha}{1+\rho\frac{1}{n-1}}, \qquad b_n = \frac{\beta+\delta\frac{1}{n-1}}{1+\rho\frac{1}{n-1}}.$$

Because b_n is known and depends on n and involves three different structural parameters, we can recover β, δ, ρ from $b_{n_1}, b_{n_2}, b_{n_3}$ if we observe three (or more) different network sizes n_1, n_2, n_3.

They are uniquely recoverable provided that $\lambda \neq 0$, where λ is the same parameter as before.

Network effects

The above model does little to accommodate correlated effects.

When at the network level, such effects are network fixed effects.

Again take

$$\boldsymbol{y} = \alpha \mathbf{1}_n + \rho \boldsymbol{P}\boldsymbol{y} + \beta \boldsymbol{x} + \delta \boldsymbol{P}\boldsymbol{x} + \boldsymbol{\varepsilon}$$

but see α as a fixed effect for the network in question (se we are sampling many small networks here).

Can difference within the network:

$$(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \rho(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P}\boldsymbol{y} + \beta(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x} + \delta(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P}\boldsymbol{x} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\varepsilon}.$$

Reduced form then comes from inverting

$$(\boldsymbol{I} - \rho\boldsymbol{P})(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \beta(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x} + \delta(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P}\boldsymbol{x} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\varepsilon}.$$

Now need that $\boldsymbol{I}_n, \boldsymbol{P}, \boldsymbol{P}^2, \boldsymbol{P}^3$ are linearly independent.