(NONLINEAR) MODELS FOR PANEL DATA

Koen Jochmans

University of Cambridge

PhD30

Last revised on January 15, 2018

Aims and scope

Aim
An overview of the large (and active) literature on panel data.

Scope
Focus will be on fixed-effect models (as opposed to random effects).

Chief difficulty: incidental-parameter problem.

Discuss a range of identification and estimation strategies for both linear and nonlinear models.

Duration
About 9 hours (hopefully).

The setting and a simple example

The structure of panel data

Units $i = 1, \ldots, N$.

Multiple observations per unit,

$$z_i = (y_i, x_i), \qquad z_i = (z_{i1}, \ldots, z_{iT});$$

the second dimension is often time, but it need not be (for example, with linked data or in network environments).

Independence in the cross section (through random sampling) is often reasonable.

More difficult in the longitudinal dimension, because we follow the same units.

Fixed effects are considered to (i) control for correlation; and (ii) to allow for additional cross-sectional heterogeneity.

We can consider three types of asymptotics:

- $N \to \infty$ with T fixed (classical asymptotics),

- $T \to \infty$ with N fixed (time-series asymptotics),

- $N \to \infty$ and $T \to \infty$ jointly (double asymptotics).

Many normal means (Neyman and Scott 1948)

We have

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

Corresponds to the simple model

$$y_{it} = \alpha_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \text{i.i.d } \mathcal{N}(0, \theta),$$

with intercepts heterogeneous across units.

An extension is a regression model (see below).

Sometimes also called error-component model, as

$$\text{var}(y_{it}) = \text{var}(E[y_{it}|\alpha_i]) + E[\text{var}(y_{it}|\alpha_i)] = \text{var}(\alpha_i) + \theta,$$

which is a within-between decomposition of the total variance.

Maximum likelihood

Random sample $\{y_{i1}, \ldots, y_{iT}\}_{i=1}^{N}$. The log-likelihood is

$$-\frac{NT}{2}\log\theta - \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{(y_{it}-\alpha_i)^2}{\theta}.$$

The first-order conditions are

$$\begin{aligned} \frac{\sum_{t=1}^{T}(y_{it}-\alpha_i)}{\theta} &= 0, \qquad \text{(for all } i\text{)}, \\ \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\alpha_i)^2}{\theta} - NT &= 0. \end{aligned}$$

The estimators are

$$\begin{aligned} \widehat{\alpha}_i &= \frac{1}{T}\sum_{t=1}^{T} y_{it} = \overline{y}_i \qquad \text{(for all } i\text{)}, \\ \widehat{\theta} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\overline{y}_i)^2. \end{aligned}$$

Classical asymptotics

Note that

$$\widehat{\alpha}_i \sim \mathcal{N}\left(\alpha_i, \theta/T\right).$$

Under classical asymptotics ($N \to \infty$, T fixed), this estimator does not converge.

Also,

$$\begin{aligned}
\widehat{\theta} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\widehat{\alpha}_i)^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left((y_{it}-\alpha_i)-(\widehat{\alpha}_i-\alpha_i)\right)^2 \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{it}^2 + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\overline{\varepsilon}_i^2 - \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{it}\,\overline{\varepsilon}_i \\
&\xrightarrow{p} \theta + \frac{\theta}{T} - 2\frac{\theta}{T} = \xrightarrow{p} \theta - \frac{\theta}{T} = \frac{T-1}{T}\theta \neq \theta
\end{aligned}$$

MLE does not do degrees-of-freedom correction for estimating α_i. Indeed,

$$\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\widehat{\alpha}_i)^2$$

would be consistent.

Distribution of α_i

Can we learn functionals of the α_i, such as $\mu = E[\alpha_i]$ or $\sigma^2 = \text{var}(\alpha_i)$?

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\alpha}_i = \frac{1}{N}\sum_{i=1}^{N}\alpha_i + \frac{1}{N}\sum_{i=1}^{N}\overline{\varepsilon}_i \xrightarrow{p} \mu$$

(because $\widehat{\alpha}_i$ is unbiased and μ is a linear functional).

However,

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\widehat{\alpha}_i^2 - \widehat{\mu}^2 = \frac{1}{N}\sum_{i=1}^{N}\alpha_i^2 - \mu^2 + \frac{1}{N}\sum_{i=1}^{N}\overline{\varepsilon}_i^2 + \frac{1}{N}\sum_{i=1}^{N}\alpha_i\,\overline{\varepsilon}_i \xrightarrow{p} \sigma^2 + \frac{\theta}{T},$$

so nonlinear functionals are estimated inconsistently.

Double asymptotics

The inconsistency of $\widehat{\theta}$ is like a bias problem in a short time series.

As $N, T \to \infty$,

$$\widehat{\alpha}_i \sim \mathcal{N}\left(\alpha_i, \frac{\theta}{T}\right) \xrightarrow{p} \alpha_i, \qquad \widehat{\theta} = \theta - \frac{\theta}{T} + O_p\left(\frac{1}{\sqrt{NT}}\right) \xrightarrow{p} \theta,$$

so all parameters can be estimated consistently.

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

$$\sqrt{NT}(\widehat{\theta} - \theta) \approx -\sqrt{\frac{N}{T}}\,\theta + O_p\left(1\right) \xrightarrow{d} \mathcal{N}\left(-c\theta, 2\theta^2\right),$$

so confidence bounds are incorrectly centered unless T grows faster than N.

Numerical example

$\theta = 1$

		BIAS		CI (classical)		CI (double)	
N	T	MLE	BC	MLE	BC	MLE	BC
100	2	−.500	.000	.003	.941	.000	.831
100	4	−.251	.000	.066	.948	.021	.901
100	8	−.125	.000	.299	.948	.220	.931
100	16	−.063	.000	.566	.948	.520	.941
10	10	−.099	.001	.847	.943	.809	.928
25	25	−.040	.000	.874	.947	.859	.943
50	50	−.012	.000	.883	.952	.875	.946
100	100	−.010	.000	.890	.952	.886	.950

$$\text{MSE} = \text{BIAS}^2 + \text{VAR} = O\left(\frac{1}{T^2}\right) + O\left(\frac{1}{NT}\right)$$

Bias is non-negligible unless $N << T$.

Distribution of α_i

Can we learn functionals of the α_i under double asymptotics, such as $\mu = E[\alpha_i]$ or $\sigma^2 = \text{var}(\alpha_i)$?

Plug-in estimator of μ is

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\alpha}_i = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{it}.$$

The second term is small compared to the first.

Under double asymptotics,

$$\sqrt{N}(\widehat{\mu} - \mu) = \sqrt{N}(\overline{\alpha} - \mu) + O_p\left(\frac{1}{\sqrt{T}}\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

has a non-degenerate limit distribution.

Conclusions

Classical asymptotics
-MLE of common parameters need not be consistent.

-Distribution of fixed effects (heterogeneity) is not estimated consistently.

Double asymptotics
-MLE and FE is consistent; but

-Distribution of MLE is incorrectly centered (bias and incorrect inference), and

-Averages across α_i converge at the slow rate of $\frac{1}{\sqrt{N}}$

Linear regression (Mundlak 1961)

Estimate agricultural production functions.

Control for soil/managerial quality (unobserved to the econometrician).

(log-linearized) Cobb-Douglas production function of farm i during season t is

$$y_{it} = x_{it}'\theta + u_{it}, \qquad u_{it} = \alpha_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \text{i.i.d. } \mathcal{N}(0, \tau^2),$$

where

- y_{it} is log-output;

- x_{it} are logged inputs (labour, capital, etc);

- α_i is an unbserved factor that remains constant across seasons (soil quality)

- ε_{it} is a random shock (rainfall and other weather conditions).

Typically, # farms >> # seasons.

Note: Key parameter is regression slope θ. Pooled least-squares estimation is inconsistent if α_i correlates with x_{it}.

So, here, the α_i and τ are nuisance parameters.

Concentrated likelihood

The joint likelihood for all parameters (θ, $\alpha_1, \ldots, \alpha_N$, and τ) is

$$-\frac{NT}{2}\log\tau^2 - \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\left(y_{it}-\alpha_i-x_{it}\theta\right)^2}{\tau^2}.$$

We concentrate the problem by replacing $\alpha_1, \ldots, \alpha_N$ and τ^2 by their maximum-likelihood estimates given θ:

$$\widehat{\alpha}_i(\theta) = \overline{y}_i - \overline{x}_i'\theta, \qquad \widehat{\tau}^2(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left((y_{it}-\overline{y}_i)-(x_{it}-\overline{x}_i)'\theta\right)^2,$$

to get the concentrated likelihood

$$-\frac{NT}{2}\log\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left((y_{it}-\overline{y}_i)-(x_{it}-\overline{x}_i)'\theta\right)^2\right) - \frac{NT}{2}.$$

Maximum likelihood

The first-order condition of the concentrated likelihood is (proportional to)

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\overline{x}_i)\left((y_{it}-\overline{y}_i)-(x_{it}-\overline{x}_i)'\theta\right)=0,$$

and yields

$$\widehat{\theta}=\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\overline{x}_i)(x_{it}-\overline{x}_i)'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\overline{x}_i)(y_{it}-\overline{y}_i)\right);$$

the within-group (least-squares) estimator.

This estimator is consistent under classical asymptotics as long as

$$E[\varepsilon_{it}|x_{i1},\ldots,x_{iT},\alpha_i]=0,$$

that is, that the regressors are strictly exogenous.

(Note that the remaining parameters are all still inconsistently estimated).

Moment interpretation

Recall the specification

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

Now,

$$E[y_{it} - x_{it}'\theta | x_{i1}, \dots, x_{iT}, \alpha_i] = \alpha_i, \qquad \text{for all } t = 1, \dots, T,$$

and so, also,

$$E[\overline{y}_i - \overline{x}_i'\theta | x_{i1}, \dots, x_{iT}, \alpha_i] = \alpha_i.$$

Hence,

$$E\left[(y_{it} - \overline{y}_i) - (x_{it} - \overline{x}_i)'\theta \middle| x_{i1}, \dots, x_{iT} \right] = 0.$$

This implies unbiasedness of the score equation for θ, and so consistency of $\widehat{\theta}$.

This result is robust to non-normality.

This is applying pooled least squares to

$$(y_{it} - \overline{y}_i) = (x_{it} - \overline{x}_i)'\theta + (\varepsilon_{it} - \overline{\varepsilon}_i),$$

the time de-meaned equations.

First differencing

Alternatively, we can first-difference the equation to get

$$(y_{it} - y_{it-1}) = (x_{it} - x_{it-1})'\theta + (\varepsilon_{it} - \varepsilon_{it-1}).$$

This gives $T - 1$ moment equations of the form

$$E[(y_{it} - y_{it-1}) - (x_{it} - x_{it-1})'\theta | x_{i1}, \ldots, x_{iT}] = 0.$$

To construct a GMM estimator, let

$$D = (T-1) \times T = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$$

and stack the $T - 1$ moments in a vector to get

$$E[D(y_i - X_i\theta)|X_i] = 0.$$

First-differenced OLS

OLS on the pooled data in first-differences solves

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

and gives the estimator

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

First-differenced GLS

Differencing induces dependence across the conditional moments.

The optimal unconditional moment condition is

$$-\sum_{i=1}^{N} X_i' D' \left(DD'\right)^{-1} D(y_i - X_i\theta) = 0$$

and yields the GLS estimator

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

A calculation gives

$$D' \left(DD'\right)^{-1} D = I - \frac{\iota\iota'}{T},$$

which is a de-meaning operator, and the GLS estimator equals the MLE.

Conclusions

In the linear regression model

Some interesting parameters can be recovered under classical asymptotics while controlling for unobserved heterogeneity.

Moment conditions are free of fixed effects and unbiased.

In general

The linear model is an exception rather than the rule.

Joint estimation of the α_i and θ will not work, in general.

Incidental-parameter problem is the failure to separate estimation of θ from estimation of α_i, in general.

We need to look for alternative estimation techniques to maximum likelihood:

-Moment conditions

-Conditional likelihood

(Approximate) likelihood separation

Adjusted likelihood and bias correction

The incidental-parameter problem and further examples

Incidental-parameter problem

Model has $y_{it}|x_i;\alpha_i,\theta \sim f(y_{it}|x_i;\alpha_i,\theta)$ and concentrated likelihood

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\log f(y_{it}|x_i;\widehat{\alpha}_i(\theta),\theta)$$

for

$$\widehat{\alpha}_i(\theta) = \arg\max_{\alpha} \frac{1}{T}\sum_{t=1}^{T}\log f(y_{it}|x_i;\alpha,\theta).$$

An infeasible estimator would use

$$\alpha_i(\theta) = \arg\max_{\alpha} E\left[\log f(y_{it}|x_i;\alpha,\theta)\right],$$

and would be consistent.

Under classical asymptotics, $\widehat{\alpha}_i(\theta)$ stays random, and $\text{plim}_{N\to\infty}\widehat{\alpha}_i(\theta) \neq \alpha_i(\theta)$.

From this, it follows that $\text{plim}_{N\to\infty}\widehat{\theta} \neq \theta$, in general.

Binary choice Logit: Chamberlain (1980)

Two-period logit model:

$$P(y_{i1}=1|x_i,\alpha_i)=F(\alpha_i), \qquad P(y_{i2}=1|x_i,\alpha_i)=F(\alpha_i+\theta),$$

for

$$F(a)=\frac{1}{1+e^{-a}}.$$

The log-likelihood is

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

Four types of individuals:

$$\begin{cases} (1,0) \text{ and } (0,1) & \text{movers out/in the sample,} \\ (1,1) \text{ and } (0,0) & \text{stayers.} \end{cases}$$

Maximum likelihood

Using that

$$f(a) = F(a)\,[1 - F(a)]$$

for logit, the score equation for α_i (given θ) is

$$y_{i1} - F(\alpha_i) + y_{i2} - F(\alpha_i + \theta) = \begin{cases} 2 - F(\alpha_i) - F(\alpha_i + \theta) & \text{if } (y_{i1}, y_{i2}) = (1,1) \\ -F(\alpha_i) - F(\alpha_i + \theta) & \text{if } (y_{i1}, y_{i2}) = (0,0) \\ 1 - F(\alpha_i) - F(\alpha_i + \theta) & \text{if } (y_{i1}, y_{i2}) = (1,0) \\ 1 - F(\alpha_i) - F(\alpha_i + \theta) & \text{if } (y_{i1}, y_{i2}) = (0,1) \end{cases} .$$

Hence,

$$\widehat{\alpha}_i(\theta) = \begin{cases} +\infty & \text{if } (y_{i1}, y_{i2}) = (1,1) \\ -\infty & \text{if } (y_{i1}, y_{i2}) = (0,0) \\ -\frac{\theta}{2} & \text{if } (y_{i1}, y_{i2}) = (1,0) \text{ or } (0,1) \end{cases} .$$

Stayers provide no information on θ. Only movers contribute to the likelihood.

Concentrated likelihood

The profile likelihood for θ is

$$\begin{aligned}\sum_{i=1}^{N} & 1\{\Delta y_i = -1\}\left\{\log F\left(-\frac{\theta}{2}\right) + \log\left(1 - F\left(\frac{\theta}{2}\right)\right)\right\} \\ & +1\{\Delta y_i = \quad 1\}\left\{\log\left(1 - F\left(-\frac{\theta}{2}\right)\right) + \log F\left(\frac{\theta}{2}\right)\right\},\end{aligned}$$

and the score is

$$\begin{aligned}& \sum_{i=1}^{N} \frac{1\{\Delta y_i = -1\}}{2}\left\{-\frac{f(-\theta/2)}{F(-\theta/2)} - \frac{f(\theta/2)}{1 - F(\theta/2)}\right\} \\ & \quad + \frac{1\{\Delta y_i = \quad 1\}}{2}\left\{\frac{f(-\theta/2)}{1 - F(-\theta/2)} + \frac{f(\theta/2)}{F(\theta/2)}\right\}, \\ & = \sum_{i=1}^{N} 1\{\Delta y_i = \quad 1\}(1 - F(\theta/2)) - 1\{\Delta y_i = -1\}\, F(\theta/2)\end{aligned}$$

Maximum likelihood estimator

Let

$$n_{01} = \sum_i 1\{\Delta y_i = -1\}, \qquad n_{10} = \sum_i 1\{\Delta y_i = 1\}.$$

The maximum likelihood estimator then satisfies

$$(n_{01} + n_{10}) F(\widehat{\theta}/2) - n_{10} = 0,$$

so that

$$\frac{\widehat{\theta}}{2} = F^{-1}\left(\frac{n_{10}}{n_{10} + n_{01}}\right) = F^{-1}\left(\frac{1}{1 + n_{01}/n_{10}}\right) \xrightarrow{p} F^{-1}\left(\frac{1}{1 + e^{-\theta}}\right).$$

Thus,

$$\widehat{\theta} \xrightarrow{p} 2F^{-1}F(\theta) = 2\theta,$$

independent of the true α_i.

Linear model: Nickell (1981)

A simple first-order autoregression has

$$y_{it} = \alpha_i + \theta y_{it-1} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}(0, \tau^2).$$

Concentrating-out the α_i gives first-order condition for θ as

$$\sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it-1} - \overline{y}_{i-}) \left((y_{it} - \overline{y}_i) - \theta (y_{it-1} - \overline{y}_{i-}) \right) = 0,$$

and the within-group least squares estimator.

However, this moment condition is not unbiased, as

$$E\left[(y_{it-1} - \overline{y}_{i-}) \left(\varepsilon_{it-1} - \overline{\varepsilon}_{i-} \right) \right] = \tau^2 N \sum_{t=1}^{T-1} \frac{T-t}{T-1} \, \theta^{t-1}.$$

y_{it} is a discounted sum of $\varepsilon_{it}, \varepsilon_{it-1}, \ldots, \varepsilon_{i1}$ and an initial condition y_{i0}. The time de-meaning introduces correlation, akin to the usual Hurwicz bias in time series.

Numerical example

An expansion (in T) of the probability limit (in N) gives

$$\text{plim}_{N\to\infty}\widehat{\theta} - \theta = -\frac{1+\theta}{T} + O\left(\frac{1}{T^2}\right).$$

		BIAS		CI	
N	T	MLE	GMM	MLE	GMM
100	4	−.413	−.054	.000	.923
100	6	−.278	−.047	.000	.910
100	8	−.206	−.039	.000	.910
100	12	−.134	−.031	.001	.900
20	20	−.081	−.089	.595	.613
50	50	−.031	−.033	.592	.603
100	100	−.015	−.016	.596	.605

Information and identification

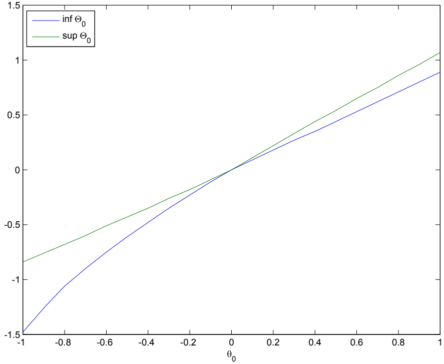
Identification

From a small-T perspective, it is not clear that parameters are point identified.

Chamberlain (2010) for binary choice:

When x_{it} has bounded support, θ is identified only in the logit model.

Autoregressive Probit model for $T = 2$, with $P(y_{i0} = 1|\alpha_i) = 1/2$ and α_i approx. normal (discretized). Identified set for θ (Honoré and Tamer 2006):

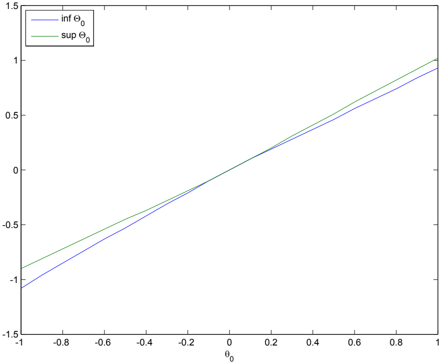


Identification (cont'd)

Information accumulates as $T \to \infty$.

Autoregressive Logit model for $T = 2$, with $P(y_{i0} = 1|\alpha_i) = 1/2$ and α_i approx. normal (discretized).

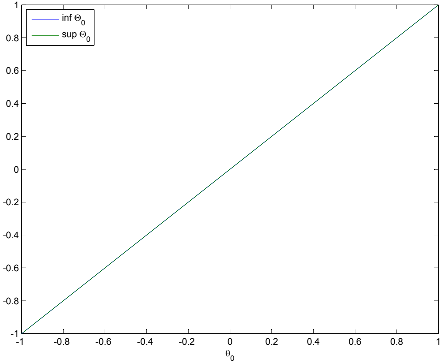
Identified set for θ:



Identification (cont'd)

Now set $T = 3$.

Identified set for θ is a singleton:



Information loss

The difficulties with binary-choice models should not be so surprising; rather the opposite, for the logit model.

Unrestricted distribution of (α_i, x_i) gets filtered to sequences of 0 and 1.

Contrary to, say, the linear model, we cannot solve for α_i.

An alternative binary-choice model (Wooldridge 1999) has

$$P[y_{it} = 1 | x_i, \alpha_i] = \alpha_i \, F(x_{it}'\theta).$$

For example,

$$y_{it} = \begin{cases} 1 & \text{if } x_{it}'\theta_0 \geq u_{it} \text{ and } \eta_i \geq v_{it} \\ 0 & \text{otherwise} \end{cases}$$

with u_{it}, v_{it} multually independent and independent of x_i, η_i.

Then

$$P[y_{it} = 1 | x_i, \eta_i] = P[x_{it}'\theta_0 \geq u_{it}, \eta_i \geq v_{it}] = F(x_{it}'\theta_0) \, G(\eta_i),$$

for $v_{it} \sim G$. Set $\alpha_i = G(\eta_i)$.

Information loss (cont'd)

The motivation for such a model is different, but so is its identifying content.

Say, for example, that

- y_{it} is decision of firm i to export to market t;
- x_{it} are market and firm characteristics determining the utility of exporting;
- η_i is the firm's type: higher η_i would be a firm that is more open toward trade.

We have

$$E[y_{it}|x_i, \alpha_i] = \alpha_i F(x_{it}'\theta_0),$$

so

$$E\left[\frac{y_{it}}{F(x_{it}'\theta_0)}\middle| x_i, \alpha_i\right] = \alpha_i, \quad \text{for all } t,$$

and, for example,

$$E\left[\frac{y_{it}}{F(x_{it}'\theta_0)} - \frac{\overline{y}_i}{\overline{F(x_i'\theta_0)}}\middle| x_i\right] = 0.$$

This is a multiplicative analog of the within-group approach in the linear model.

Conditional likelihood

Sufficient statistics

Generic setting: $y_{it}|x_i, \alpha_i$ has density $f(y_{it}|x_i, \alpha_i; \theta)$. Likelihood contribution of i is

$$f(y_i|x_i; \alpha_i, \theta) = \prod_{t=1}^{T} f(y_{it}|x_i; \alpha_i, \theta).$$

Let $s_i = s(y_i)$ for some function s; s_i has density $f(s_i|x_i; \alpha_i, \theta)$

Then

$$f(y_i|s_i, x_i; \alpha_i, \theta) = \frac{f(y_i, s_i|x_i; \alpha_i, \theta)}{f(s_i|x_i; \alpha_i, \theta)} = \frac{f(y_i, |x_i; \alpha_i, \theta)}{f(s_i|x_i; \alpha_i, \theta)}.$$

We call s_i a sufficient statistic for α_i if

$$f(y_i|x_i, s_i; \alpha_i, \theta) = f(y_i|x_i, s_i; \theta);$$

given s_i, the distribution of y_i does no longer depend on α_i. So we have separated estimation of α_i from inference on θ.

We can then maximize the conditional log-likelihood

$$\sum_{i=1}^{N} \log f(y_i|x_i, s_i; \theta).$$

Sufficient statistics: Linear model

Scope of the conditional-likelihood approach is limited by the existence of a sufficient statistic (which is not guaranteed).

Sufficient statistics are found by trail and error.

Consider the linear model

$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathcal{N}(0, \sigma^2),$$

and $\theta = (\beta, \sigma^2)$.

A sufficient statistic here is

$$s_i = T^{-1} \sum_{t=1}^{T} y_{it} = \overline{y}_i$$

(or any one-to-one transformation of it).

Linear model

First, because

$$y_i|X_i;\alpha_i,\beta,\sigma^2 \sim \mathcal{N}(\alpha_i\iota_T + X_i\theta, \sigma^2 I),$$

we have

$$f(y_i|X_i;\alpha_i,\beta,\sigma^2) = (2\pi\sigma^2)^{-T/2} \exp\left(-\frac{\sum_{t=1}^{T}(y_{it}-\alpha_i-x'_{it}\beta)^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-T/2} \exp\left(-\frac{\sum_{t=1}^{T}((y_{it}-\overline{y}_i)-(x_{it}-\overline{x}_i)'\beta)^2 + T(\overline{y}_i-\alpha_i-\overline{x}'_i\beta)^2}{2\sigma^2}\right).$$

Linear model (cont'd)

Second,

$$s_i|X_i;\alpha_i,\beta,\sigma^2 \sim \mathcal{N}\left(\alpha_i + \overline{x}_i'\beta, \frac{\sigma^2}{T}\right).$$

So

$$f(s_i|X_i;\alpha_i,\beta,\sigma^2) = (2\pi\sigma^2/T)^{-1/2}\exp\left(-\frac{(\overline{y}_i - \alpha_i - \overline{x}_i'\beta)^2}{2\sigma^2/T}\right).$$

Put together we get

$$f(y_i|X_i,s_i,\alpha_i,\beta,\sigma^2) = (2\pi\sigma^2)^{-(T-1)/2}\exp\left(-\frac{\sum_{t=1}^{T}((y_{it} - \overline{y}_i) - (x_{it} - \overline{x}_i)'\beta)^2}{2\sigma^2}\right).$$

The conditional log-likelihood then is

$$\frac{N(T-1)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\left((y_{it} - \overline{y}_{i\cdot}) - (x_{it} - \overline{x}_{i\cdot})'\beta\right)^2.$$

The second term is the within-group objective function, the first term leads to the degrees-of-freedom correction for $\widehat{\sigma}^2$.

Two-period logit

As before,

$$P[y_{it} = 1 | x_i, \alpha_i] = F(\alpha_i + x_{it}'\theta), \qquad F(a) = \frac{1}{1+e^{-a}}.$$

Set $T = 2$. Then we have four possible sequences, with probabilities

$$\begin{aligned}
P[y_{i1} = 0, y_{i2} = 0 | x_i, \alpha_i] &= [1 - F(\alpha_i + x_{i1}'\theta)]\,[1 - F(\alpha_i + x_{i2}'\theta)] \\
P[y_{i1} = 1, y_{i2} = 1 | x_i, \alpha_i] &= F(\alpha_i + x_{i1}'\theta)\,F(\alpha_i + x_{i2}'\theta) \\
P[y_{i1} = 0, y_{i2} = 1 | x_i, \alpha_i] &= [1 - F(\alpha_i + x_{i1}'\theta)]\,F(\alpha_i + x_{i2}'\theta) \\
P[y_{i1} = 1, y_{i2} = 0 | x_i, \alpha_i] &= F(\alpha_i + x_{i1}'\theta)\,[1 - F(\alpha_i + x_{i2}'\theta)].
\end{aligned}$$

A sufficient statistic here is $s_i = y_{i1} + y_{i2}$.

Conditional probabilities

Non-movers drop out/contain no information. Indeed,

$P[y_{i1}=0,y_{i2}=0|x_i,\alpha_i,s_i=0]=1, \qquad P[y_{i1}=1,y_{i2}=1|x_i,\alpha_i,s_i=2]=1.$

The only remaining cases are the mover probabilities:

$$P[y_{i1}=1,y_{i2}=0|x_i,\alpha_i,s_i=1], \qquad P[y_{i1}=0,y_{i2}=1|x_i,\alpha_i,s_i=1],$$

which form a proper probability distribution.

$$\begin{aligned} P[s_i=1|x_i,\alpha_i] &= P[y_{i1}=0,y_{i2}=1|x_i,\alpha_i]+P[y_{i1}=1,y_{i2}=0|x_i,\alpha_i] \\ &= \frac{e^{-\alpha_i - x'_{i1}\theta}}{(1+e^{-\alpha_i - x'_{i1}\theta})(1+e^{-\alpha_i - x'_{i2}\theta})} + \frac{e^{-\alpha_i - x'_{i2}\theta}}{(1+e^{-\alpha_i - x'_{i1}\theta})(1+e^{-\alpha_i - x'_{i2}\theta})} \end{aligned}$$

So,

$$\begin{aligned} P[y_{i1}=1,y_{i2}=0|x_i,\alpha_i,s_i=1] &= \frac{e^{-\alpha_i - x'_{i2}\theta}}{e^{-\alpha_i - x'_{i1}\theta}+e^{-\alpha_i - x'_{i2}\theta}} \\ &= \frac{1}{1+e^{(x_{i2}-x_{i1})'\theta}} = F(-(x_{i2}-x_{i1})'\theta) \end{aligned}$$

Conditional likelihood

And, also,

$$P[y_{i1} = 0, y_{i2} = 1 | x_i, \alpha_i, s_i = 1] = 1 - F(-(x_{i2} - x_{i1})'\theta) = F((x_{i2} - x_{i1})'\theta).$$

The conditional log-likelihood thus is

$$\sum_{i=1}^{N} 1\{\Delta y_i = 1\} \log F(\Delta x_i'\theta) + 1\{\Delta y_i = -1\} \log(1 - F(\Delta x_i'\theta))$$

This is a logit likelihood for first-differenced outcomes $\Delta y_i = y_{i2} - y_{i1}$ in the subpopulation of movers.

Indeed,

$$\Delta y_i | x_i, s_i = 1 \sim \text{ Bernoulli with } F(\Delta x_i'\theta_0).$$

Conditional moment conditions

The first-order condition of conditional likelihood is

$$\sum_{i=1}^{N} x_i \left[1\{\Delta y_i = 1\} - F(\Delta x_i' \theta)\right] 1\{\Delta y_i \neq 0\}.$$

This is a particular version of the conditional moment condition

$$E[1\{\Delta y_i = 1\} - F(\Delta x_i' \theta) | x_i, \Delta y_i \neq 0] = 0,$$

giving a GMM interpretation to conditional logit.

Miracle of the logit comes from functional form.

For example: No sufficient statistics for Probit model.

In fact, no point identification except in logit, in general.

Autoregressive logit

Purely dynamic model with

$$P(y_{it} = |y_{it-1}, \ldots, y_{i0}, \alpha_i) = F(\alpha_i + \theta y_{it-1})$$

and unrestricted initial condition; i.e., $P(y_{i0} = 1|\alpha_i) = p_i$.

With the initial condition unrestricted, p_i essentially acts as another fixed effect.

Point identification fails for $T = 2$ (see above) and $\widehat{\theta} = \pm\infty$ in this case.

Set $T = 3$; so 3 effective observations and 1 initial condition. Consider sequences where

$$y_{i1} + y_{i2} = 1.$$

We will get information out of sequences for which

$$\{y_{i0} + y_{i3} = 1, y_{i1} + y_{i2} = 1\}.$$

Four such sequences:

$$(1,1,0,0) \quad \text{and} \quad (1,0,1,0); \quad \text{and} \quad (0,1,0,1) \quad \text{and} \quad (0,0,1,1).$$

Choice probabilities

We have

$$P(y_{i0} = y_0, y_{i1} = 0, y_{i2} = 1, y_{i3} = y_3 | \alpha_i) =$$
$$p_i^{y_0}\,(1 - p_i)^{1-y_0}\,[1 - F(\alpha_i + \theta y_0)]\,F(\alpha_i)\,F(\alpha + \theta)^{y_3}\,[1 - F(\alpha_i + \theta)]^{1-y_3}$$

$$P(y_{i0} = y_0, y_{i1} = 1, y_{i2} = 0, y_{i3} = y_3 | \alpha_i) =$$
$$p_i^{y_0}\,(1 - p_i)^{1-y_0}\,F(\alpha_i + \theta y_0)\,[1 - F(\alpha_i + \theta)]\,F(\alpha)^{y_3}\,[1 - F(\alpha_i)]^{1-y_3}$$

So,

$$\frac{P(y_{i0} = y_0, y_{i1} = 1, y_{i2} = 0, y_{i3} = y_3 | \alpha_i)}{P(y_{i0} = y_0, y_{i1} = 0, y_{i2} = 1, y_{i3} = y_3 | \alpha_i)} =$$
$$\frac{[1 - F(\alpha_i + \theta y_0)]\,F(\alpha_i)\,F(\alpha_i + \theta)^{y_3}\,[1 - F(\alpha_i + \theta)]^{1-y_3}}{F(\alpha_i + \theta y_0)\,[1 - F(\alpha_i + \theta)]\,F(\alpha_i)^{y_3}\,[1 - F(\alpha_i)]^{1-y_3}} =$$
$$\begin{cases} 1 & \text{if } y_0 = y_3 \\ \frac{[1-F(\alpha_i)]\,F(\alpha_i+\theta)}{F(\alpha_i)\,[1-F(\alpha_i+\theta)]} & \text{if } y_0 < y_3 \\ \frac{[1-F(\alpha_i+\theta)]\,F(\alpha_i)}{F(\alpha_i+\theta)\,[1-F(\alpha_i)]} & \text{if } y_0 > y_3 \end{cases}$$

Choice probabilities for $y_0 > y_3$

First note that

$$P(y_{i1}=1, y_{i2}=0, y_{i3}=0|y_{i0}=1,\alpha_i) = F(\alpha_i+\theta)\left[1-F(\alpha_i+\theta)\right]\left[1-F(\alpha_i)\right]$$

and

$$P(y_{i1}=0, y_{i2}=1, y_{i3}=0|y_{i0}=1,\alpha_i) = \left[1-F(\alpha_i+\theta)\right]F(\alpha_i)\left[1-F(\alpha_i+\theta)\right]$$

Thus,

$$\begin{aligned}\frac{P(y_{i1}=0, y_{i2}=1, y_{i3}=0|y_{i0}=1,\alpha_i)}{P(y_{i1}=1, y_{i2}=0, y_{i3}=0|y_{i0}=1,\alpha_i)} &= \frac{F(\alpha_i)\left[1-F(\alpha_i+\theta)\right]}{F(\alpha_i+\theta)\left[1-F(\alpha_i)\right]} \\ &= \frac{\frac{1}{1+e^{-\alpha_i}}\,\frac{e^{-\alpha_i-\theta}}{1+e^{-\alpha_i-\theta}}}{\frac{1}{1+e^{-\alpha_i-\theta}}\,\frac{e^{-\alpha_i}}{1+e^{-\alpha_i}}} = \frac{e^{-\alpha_i-\theta}}{e^{-\alpha_i}} = e^{-\theta}\end{aligned}$$

And so

$$\frac{P(y_{i1}=1, y_{i2}=0, y_{i3}=0|y_{i0}=1,\alpha_i)}{P(y_{i1}\neq y_{i2}, y_{i3}=0|y_{i0}=1,\alpha_i)} = \frac{1}{1+e^{-\theta}} = F(\theta)$$

$$\frac{P(y_{i1}=0, y_{i2}=1, y_{i3}=0|y_{i0}=1,\alpha_i)}{P(y_{i1}\neq y_{i2}, y_{i3}=0|y_{i0}=1,\alpha_i)} = \frac{e^{-\theta}}{1+e^{-\theta}} = 1-F(\theta)$$

Choice probabilities for $y_0 < y_3$

Similarly

$$P(y_{i1} = 1, y_{i2} = 0, y_{i3} = 1 | y_{i0} = 0, \alpha_i) = F(\alpha_i)\left[1 - F(\alpha_i + \theta)\right] F(\alpha_i)$$

and

$$P(y_{i1} = 0, y_{i2} = 1, y_{i3} = 1 | y_{i0} = 0, \alpha_i) = \left[1 - F(\alpha_i)\right] F(\alpha_i) F(\alpha_i + \theta)$$

Thus,

$$\frac{P(y_{i1} = 1, y_{i2} = 0, y_{i3} = 1 | y_{i0} = 0, \alpha_i)}{P(y_{i1} = 0, y_{i2} = 1, y_{i3} = 1 | y_{i0} = 0, \alpha_i)} = \frac{F(\alpha_i)\left[1 - F(\alpha_i + \theta)\right]}{F(\alpha_i + \theta)\left[1 - F(\alpha_i)\right]} = \frac{e^{-\alpha_i - \theta}}{e^{-\alpha_i}} = e^{-\theta}$$

And so

$$\frac{P(y_{i1} = 0, y_{i2} = 1, y_{i3} = 1 | y_{i0} = 0, \alpha_i)}{P(y_{i1} \neq y_{i2}, y_{i3} = 1 | y_{i0} = 0, \alpha_i)} = \frac{1}{1 + e^{-\theta}} = F(\theta)$$

$$\frac{P(y_{i1} = 1, y_{i2} = 0, y_{i3} = 1 | y_{i0} = 0, \alpha_i)}{P(y_{i1} \neq y_{i2}, y_{i3} = 1 | y_{i0} = 0, \alpha_i)} = \frac{e^{-\theta}}{1 + e^{-\theta}} = 1 - F(\theta)$$

Conditional likelihood

Put together, we can write the conditional likelihood compactly as

$$\sum_{i=1}^{N} F((y_{i0}-y_{i3})\theta)^{1\{y_{i1}>y_{i2}\}}\left[1-F((y_{i0}-y_{i3})\theta)\}\right]^{1\{y_{i1}<y_{i2}\}}.$$

The implied estimating equation is

$$\sum_{i=1}^{N} 1\{y_{i0}>y_{i3}\}\left[1\{y_{i1}>y_{i2}\}-F(\theta)\right]+1\{y_{i0}<y_{i3}\}\left[1\{y_{i1}<y_{i2}\}-F(\theta)\right]=0$$

and implicitly conditions on

$$\begin{pmatrix} y_{i0}+y_{i3} \\ y_{i1}+y_{i2} \end{pmatrix}=\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Autoregressive logit with regressors

The conditional likelihood above implicitly uses the fact that probabilities are time invariant.

This is no longer true with time varying regressors, as in

$$P(y_{it} = 1 | x_i, y_{it-1}, \alpha_i) = F(\alpha_i + \rho y_{it-1} + x'_{it}\beta)$$

and $\theta = (\rho, \beta)'$.

We have

$$P(y_{i1} = 0, y_{i2} = 1, y_{i3} = y_3 | y_{i0} = y_0, x_i, \alpha_i) =$$
$$[1 - F(\alpha_i + \rho y_0 + x'_{i1}\beta)]\, F(\alpha_i + x'_{i2}\beta)\, F(\alpha + \rho + x'_{i3}\beta)^{y_3}\, [1 - F(\alpha_i + \rho + x'_{i3}\beta)]^{1-y_3}$$

$$P(y_{i1} = 1, y_{i2} = 0, y_{i3} = y_3 | y_{i0} = y_0, x_i, \alpha_i) =$$
$$F(\alpha_i + \rho y_0 + x'_{i1}\beta)\, [1 - F(\alpha_i + \rho + x'_{i2}\beta)]\, F(\alpha + x'_{i3}\beta)^{y_3}\, [1 - F(\alpha_i + x'_{i3}\beta)]^{1-y_3}$$

Conditional probabilities

The impact of α_i is related to the distance between x_{i2} and x_{i3}. For example, $P[y_{i1} > y_{i2} | y_{i1} \neq y_{i2}, y_{i0} > y_{i3}, x_i, \alpha_i]$ equals

$$\frac{[1 - F(\alpha_i + x_{i3}'\beta)] \boxed{[1 - F(\alpha_i + \rho + x_{i2}'\beta)]} F(\alpha_i + \rho + x_{i1}'\beta)}{\left\{ \begin{array}{l} [1 - F(\alpha_i + x_{i3}'\beta)] \boxed{[1 - F(\alpha_i + \rho + x_{i2}'\beta)]} F(\alpha_i + \rho + x_{i1}'\beta) \\ + \boxed{[1 - F(\alpha_i + \rho + x_{i3}'\beta)]} F(\alpha_i + x_{i2}'\beta) [1 - F(\alpha_i + \rho + x_{i1}'\beta)] \end{array} \right\}}$$

and $P[y_{i1} < y_{i2} | y_{i1} \neq y_{i2}, y_{i0} > y_{i3}, x_i, \alpha_i]$ equals

$$\frac{\boxed{[1 - F(\alpha_i + \rho + x_{i3}'\beta)]} F(\alpha_i + x_{i2}'\beta) [1 - F(\alpha_i + \rho + x_{i1}'\beta)]}{\left\{ \begin{array}{l} [1 - F(\alpha_i + x_{i3}'\beta)] \boxed{[1 - F(\alpha_i + \rho + x_{i2}'\beta)]} F(\alpha_i + \rho + x_{i1}'\beta) \\ + \boxed{[1 - F(\alpha_i + \rho + x_{i3}'\beta)]} F(\alpha_i + x_{i2}'\beta) [1 - F(\alpha_i + \rho + x_{i1}'\beta)] \end{array} \right\}}.$$

The crucial difference with before is the discrepancy between the boxed terms.

Conditional probabilities with $x_{i2} = x_{i3}$

When $x_{i2} = x_{i3}$ we get

$$\frac{P[y_{i1} > y_{i2} | y_{i1} \neq y_{i2}, y_{i0} > y_{i3}, x_i, \alpha_i]}{P[y_{i1} < y_{i2} | y_{i1} \neq y_{i2}, y_{i0} > y_{i3}, x_i, \alpha_i]} = e^{\rho - (x_{i2} - x_{i1})'\beta}$$
$$\frac{P[y_{i1} > y_{i2} | y_{i1} \neq y_{i2}, y_{i0} < y_{i3}, x_i, \alpha_i]}{P[y_{i1} < y_{i2} | y_{i1} \neq y_{i2}, y_{i0} < y_{i3}, x_i, \alpha_i]} = e^{-\rho - (x_{i2} - x_{i1})'\beta}$$
$$\frac{P[y_{i1} > y_{i2} | y_{i1} \neq y_{i2}, y_{i0} = y_{i3}, x_i, \alpha_i]}{P[y_{i1} < y_{i2} | y_{i1} \neq y_{i2}, y_{i0} = y_{i3}, x_i, \alpha_i]} = e^{-(x_{i2} - x_{i1})'\beta}.$$

We thus have conditional probabilities

$$P[y_{i1} < y_{i2} | y_{i1} \neq y_{i2}, y_{i0} \neq y_{i3}, x_{i2} = x_{i3}] = F((y_{i0} - y_{i3})\rho + (x_{i2} - x_{i1})'\beta)$$

and

$$P[y_{i1} > y_{i2} | y_{i1} \neq y_{i2}, y_{i0} \neq y_{i3}, x_{i2} = x_{i3}] = 1 - F((y_{i0} - y_{i3})\rho + (x_{i2} - x_{i1})'\beta)$$

that are free of fixed effects.

Note that the events with $y_{i0} = y_{i3}$ are not informative about ρ but they are informative about β.

Conditional likelihood

The conditional likelihood becomes

$$\sum_{i=1}^{N} 1\{y_{i1} \neq y_{i3}\}\, k(x_{i2} - x_{i3})\, 1\{y_{i2} > y_{i1}\} \log F((y_{i0} - y_{i3})\rho + (x_{i2} - x_{i1})'\beta)$$
$$+ \sum_{i=1}^{N} 1\{y_{i1} \neq y_{i3}\}\, k(x_{i2} - x_{i3})\, 1\{y_{i2} < y_{i1}\} \log(1 - F((y_{i0} - y_{i3})\rho + (x_{i2} - x_{i1})'\beta))$$

When x_i is discrete we set

$$k(x_{i2} - x_{i3}) = 1\{x_{i2} - x_{i3} = 0\}$$

and the estimator has standard properties.

When x_i is continuous we smooth across observations by setting

$$k(x_{i2} - x_{i3}) = \frac{1}{h} K\left(\frac{x_{i2} - x_{i3}}{h}\right)$$

and the estimator will have a nonparametric convergence rate.

Note that, in any event, we need that $P(|x_{i2} - x_{i3}| < \epsilon) > \delta > 0$ for any small $\epsilon > 0$.

Poisson model

The Poisson model has

$$P[y_{it}|x_i, \alpha_i] = \frac{(\alpha_i \ \exp(x_{it}'\theta))^{y_{it}} \ \exp\big(-\alpha_i \exp(x_{it}'\theta)\big)}{y_{it}!}$$

With $s_i = T^{-1}\sum_{t=1}^{T} y_{it}$, we have

$$y_{i1}, \ldots, y_{iT}|s_i \sim \text{Multinomial}\left(T; \frac{\exp(x_{i1}'\theta)}{\sum_t \exp(x_{it}'\theta)}, \ldots, \frac{\exp(x_{iT}'\theta)}{\sum_t \exp(x_{it}'\theta)}\right).$$

Hence, s_i is a sufficient statistic and the conditional likelihood,

$$\sum_{i=1}^{N} \ell_i^c(\theta) = \sum_i \sum_t \frac{T!}{y_{i1}! \cdots y_{iT}!} \log\left[\frac{\exp(x_{it}'\theta)}{\sum_{t=1}^{T} \exp(x_{it}'\theta)}\right]^{y_{it}},$$

yields a consistent estimator.

Hausman et al. (1984).

Full maximum likelihood estimation

Consider the reparametrization

$$\eta_i = \alpha_i \sum_{t=1}^{T} \lambda_{it}(\theta).$$

The associated likelihood contribution of unit i is

$$e^{\ell_i(\theta,\eta_i)} \propto \exp(-\eta_i)\eta_i^{\sum_{t=1}^{T} y_{it}} \; e^{\ell_i^c(\theta)} \longrightarrow e^{\ell_i(\theta,\eta_i)} \propto e^{\ell_i^c(\theta)},$$

so we achieve complete likelihood separation.

Hence,

$$\frac{\partial \ell_i(\theta,\eta_i)}{\partial \theta} = \frac{\partial \ell_i^c(\theta)}{\partial \theta}$$

Because the MLE is invariant with respect to reparametrizations, the simple fixed-effect ML estimator already gives a consistent point estimate.

So, there are incidental parameters in the Poisson model, but there is no incidental-parameter problem (for θ).

See Lancaster (2002).

Absence of score bias

The concentrated score equation of maximum likelihood is unbiased.

A calculation gives

$$\frac{\partial \ell_i(\theta, \widehat{\alpha}_i(\theta))}{\partial \theta} = \sum_t x_{it} \left(y_{it} - \exp(x_{it}'\theta)\, \widehat{\alpha}_i(\theta) \right),$$

where

$$\widehat{\alpha}_i(\theta) = \frac{\sum_t y_{it}}{\sum_t \exp(x_{it}'\theta)}.$$

Recall that

$$E[y_{it}|x_i, \alpha_i] = \exp(x_{it}'\theta)\, \alpha_i.$$

So,

$$E[\widehat{\alpha}_i(\theta)|x_i, \alpha_i] = \frac{\sum_t E[y_{it}|x_i, \alpha_i]}{\sum_t \exp(x_{it}'\theta)} = \alpha_i,$$

and

$$E\left[\frac{\partial \ell_i(\theta, \widehat{\alpha}_i(\theta))}{\partial \theta} \middle| x_i \right] = 0.$$

Moment interpretation

We have

$$E\left[\frac{y_{it}}{\exp(x'_{it}\theta)}\,\middle|\,x_i,\alpha_i\right]=\alpha_i,\qquad\text{(for all } t\text{)},$$

and so for any function g,

$$\sum_i\sum_t g(x_{it};\theta)\left[\frac{y_{it}}{\exp(x'_{it}\theta)}-\frac{\bar{y}_i}{\overline{\exp(x'_{it}\theta)}}\right]=0$$

is an unbiased estimating equation for θ.

The score equation for maximum likelihood can be written as

$$\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}\exp(x'_{it}\theta)\left[\frac{y_{it}}{\exp(x'_{it}\theta)}-\frac{\bar{y}_i}{\overline{\exp(x'_{it}\theta)}}\right]=0,$$

and is a special case of the above moment condition, with $g(x_{it};\theta)=x_{it}\exp(x'_{it}\theta)$.

Adjusted likelihood

Bias calculations

Conditional likelihood arises from viewing incidental-parameter problem as failure to separate inference on θ from estimation of α_i.

Adjusted likelihood views incidental-parameter problem as a problem of a biased estimating equation.

The MLE maximizes the concentrated likelihood

$$\sum_i \ell_i(\theta, \widehat{\alpha}_i(\theta)).$$

Incidental-parameter bias arises because

$$E\left[\frac{\partial \ell_i(\theta, \widehat{\alpha}_i(\theta))}{\partial \theta}\right] \neq 0;$$

if not there is no problem (recall the Poisson model, for example).

We can compute the score bias under (θ, α_i) (if need be, by simulation).

Fixed-T estimation is possible if the bias is free of incidental parameters.

Many normal means

Many normal means,

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

From before,

$$\frac{\partial \ell_i(\theta, \widehat{\alpha}_i(\theta))}{\partial \theta} = \frac{\sum_{t=1}^{T} (y_{it} - \overline{y}_i)^2}{2\theta^2} - \frac{T}{2\theta}$$

and

$$E\left[\frac{\partial \ell_i(\theta, \widehat{\alpha}_i(\theta))}{\partial \theta}\right] = \frac{\sum_{t=1}^{T} E[(y_{it} - \overline{y}_i)^2]}{2\theta^2} - \frac{T}{2\theta} = -\frac{1}{2\theta}.$$

So, by construction,

$$\sum_{i=1}^{N} \frac{\partial \ell_i(\theta, \widehat{\alpha}_i(\theta))}{\partial \theta} + \frac{N}{2\theta} = 0$$

is an unbiased estimating equation. The correction factor adjust the degrees of freedom.

Because $\frac{N}{2} \partial \log \theta / \partial \theta = N/(2\theta)$, an adjusted likelihood is

$$-\frac{N(T-1)}{2} \log \theta - \frac{\sum_i \sum_t (y_{it} - \overline{y}_i)^2}{2\theta}$$

which co-incides with the conditional likelihood for this problem.

Linear autoregression

Autoregressive model

$$y_{it} = \alpha_i + \rho y_{it-1} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim /, \sigma^{\in}.$$

Here,

$$\frac{\partial \ell_i(\theta, \widehat{\alpha}_i(\theta))}{\partial \theta} = \begin{pmatrix} \frac{\partial \ell_i(\theta,\widehat{\alpha}_i(\theta))}{\partial \rho} \\ \frac{\partial \ell_i(\theta,\widehat{\alpha}_i(\theta))}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{\sum_{t=1}^{T} (y_{it-1}-\overline{y}_{i-})((y_{it}-\overline{y}_{i})-(y_{it-1}-\overline{y}_{i-})\rho)}{\sigma^2} \\ \frac{\sum_{t=1}^{T} ((y_{it}-\overline{y}_{i})-(y_{it-1}-\overline{y}_{i-})\rho)^2}{2\theta^2} - \frac{T}{2\theta} \end{pmatrix}.$$

The bias (conditional on y_{i0}) is

$$E\left[\frac{\partial \ell_i(\theta, \widehat{\alpha}_i(\theta))}{\partial \theta}\right] = \begin{pmatrix} \frac{\sum_{t=1}^{T} E[(y_{it-1}-\overline{y}_{i-})\,(\varepsilon_{it}-\overline{\varepsilon}_{i})]}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} = -\begin{pmatrix} \sum_{t=1}^{T-1} \frac{T-t}{T-1}\, \rho^{t-1} \\ \frac{1}{2\sigma^2} \end{pmatrix}.$$

Adjusted likelihood becomes

$$-\frac{N(T-1)}{2} \log\theta - \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} ((y_{it}-\overline{y}_{i})-(y_{it-1}-\overline{y}_{i-})\rho)^2}{2\theta} + N \sum_{t=1}^{T-1} \frac{T-t}{T-1}\frac{1}{t}\rho^{t}.$$

Lancaster (2002), Dhaene and Jochmans (2015).

Weibull model for duration data

Duration data with survival function

$$P(y_{it} > y|x_i, \alpha_i) = e^{-\left(\frac{y}{\alpha_i\, e^{x'_{it}\beta}}\right)^{\kappa}}$$

and hazard function

$$\frac{\kappa}{\alpha_i\, e^{x'_{it}\beta}} \left(\frac{y}{\alpha_i\, e^{x'_{it}\beta}}\right)^{\kappa-1}$$

for $y \geq 0$.

The schape parameter $\kappa > 0$ drives the shape of the hazard function; β are regression slopes that affect the scale.

The density of $y_{it}|x_i, \alpha_i$ is

$$\frac{\kappa}{\alpha_i e^{x'_{it}\beta}} \left(\frac{y_{it}}{\alpha_i e^{x'_{it}\beta}}\right)^{\kappa-1} e^{-(y_{it}/\alpha_i e^{x'_{it}\beta})^{\kappa}}$$

Alternative view:

$$y_{it}^{\kappa} \sim \operatorname{Exp} \alpha_i e^{(x'_{it}\beta)},$$

so the Weibull distribution generalizes the exponential distribution ($\kappa = 1$) to allow for non-constant hazard rates.

Concentrated likelihood and score

The concentrated likelihood is

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\log\kappa+\kappa\log y_{it}-\log\left(\sum_{t=1}^{T}w_{it}(\beta,\kappa)-\kappa x_{it}'\beta\right)$$

for

$$w_{it}(\beta,\kappa)=\left(y_{it}\,e^{-x_{it}'\beta}\right)^{\kappa}.$$

Associated score for unit i is

$$\begin{pmatrix}\frac{\partial\ell_i(\theta,\widehat{\alpha}_i(\theta))}{\partial\beta}\\ \frac{\partial\ell_i(\theta,\widehat{\alpha}_i(\theta))}{\partial\kappa}\end{pmatrix}=\begin{pmatrix}\kappa\sum_{t=1}^{T}x_{it}\left(\frac{w_{it}(\beta,\kappa)}{\sum_{t=1}^{T}w_{it}(\beta,\kappa)}-1\right)\\ \sum_{t=1}^{T}\left(\frac{1}{\kappa}+\log y_{it}-\frac{1}{\kappa}\frac{\sum_{t=1}^{T}w_{it}(\beta,\kappa)\,\log w_{it}(\beta,\kappa)}{\sum_{t=1}^{T}w_{it}(\beta,\kappa)}-x_{it}'\beta\right)\end{pmatrix}.$$

A calculation gives

$$E\left[\begin{pmatrix}\frac{\partial\ell_i(\theta,\widehat{\alpha}_i(\theta))}{\partial\beta}\\ \frac{\partial\ell_i(\theta,\widehat{\alpha}_i(\theta))}{\partial\kappa}\end{pmatrix}\right]=\begin{pmatrix}0\\ \frac{1}{\kappa}\end{pmatrix}.$$

Numerical example

β, κ are not information orthogonal, so bias is transferred.

	bias		std	
T	$\widehat{\beta}$	$\widehat{\kappa}$	$\widehat{\beta}$	$\widehat{\kappa}$
2	.7005	.7006	.2483	.1437
4	.2234	.2230	.0995	.0581
6	.1289	.1285	.0684	.0410
12	.0569	.0572	.0412	.0248

	bias		std	
T	$\dot{\beta}$	$\dot{\kappa}$	$\dot{\beta}$	$\dot{\kappa}$
2	.0124	.0104	.1458	.0848
4	.0070	.0068	.0818	.0478
6	.0023	.0019	.0607	.0363
12	.0005	.0008	.0391	.0235

Simulations: $N = 100$, $\beta = \kappa = 1$, $x_{it} \sim \mathcal{N}(0,1)$, $\alpha_i \sim \mathcal{N}(0,1)$.

Moment conditions

Linear autoregression

Yet an alternative approach is to drop the likelihood framework completely and look for moments for θ that do not involve α_i.

In the autoregression

$$y_{it} = \alpha_i + \theta y_{it-1} + \varepsilon_{it},$$

lags are valid instruments for first-differences under restrictions on serial correlation in the errors.

With

$$\Delta y_{it} = \Delta y_{it-1}\theta + \Delta\varepsilon_{it},$$

we have

$$E[\Delta y_{it} - \Delta y_{it-1}\theta | y_{it-2}, \ldots, y_{i0}] = 0$$

if $E[\varepsilon_{it} | y_{it-1}, \ldots, y_{i0}, \alpha_i] = 0$.

Absence of serial correlation in ε_{it} is enough to justify

$$E\left[\begin{pmatrix} y_{it-2} \\ y_{it-3} \\ \vdots \\ y_{i0} \end{pmatrix} \left(\Delta y_{it} - \Delta y_{it-1}\theta\right)\right] = 0, \qquad (\text{for all } t = 2, \ldots, T).$$

GMM estimation

In matrix form we have

$$E[Z_i'(\Delta y_i - \Delta y_{i-}\theta)] = 0$$

for

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

GMM with weight matrix W is available in closed form as is

$$\left(\left(\sum_i Z_i'\Delta y_{i-}\right)' W \left(\sum_i Z_i'\Delta y_{i-}\right)\right)^{-1} \left(\left(\sum_i Z_i'\Delta y_{i-}\right)' W \left(\sum_i Z_i'\Delta y_i\right)\right)^{-1};$$

one step GMM uses $W^{-1} = \sum_i Z_i'DD'Z_i$ while two-step uses $W^{-1} = \sum_i Z_i'\Delta\widehat{\varepsilon}_i\widehat{\varepsilon}_i'Z_i$

Anderson and Hsiao (1981, 1982), Holtz-Eakin, Newey, and Rosen (1988), Arellano and Bond (1991).

Identification

Be careful with this estimator when T/N is not very small (see later).

θ is globally identified if

$$\operatorname{rank} E[Z_i' \Delta y_{i-}] = T - 1.$$

In the unit root model

$$y_{it} = y_{it-1} + \varepsilon_{it}$$

the relevance condition fails because

$$E[y_{it-2-j}\,(y_{it-1} - y_{it-2})] = E[y_{it-2-j}\,\varepsilon_{it-1}] = 0$$

(for $j \geq 0$).

Similarly, in a model with high persistence, such as

$$y_{it} = (1-\theta)\,\alpha_i + \theta y_{it-1} + \varepsilon_{it},$$

the instruments are weakly correlated and so the Jacobian is nearly singular.

General predetermined regressors

GMM is equally applicable in the model

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

where, now,

$$E[\varepsilon_{it}|x_{it}, \ldots, x_{i1}, \alpha_i] = 0$$

as opposed to

$$E[\varepsilon_{it}|x_{iT}, \ldots, x_{i1}, \alpha_i] = 0$$

(as above).

So, models with feedback can be estimated in the same way.

Poisson model

Poisson maximum likelihood fails when regressors are predetermined. However, the moment approach can still be used.

Say

$$E[y_{it}|x_{i1},\ldots,x_{it},\alpha_i] = \lambda(x_{it}\theta)\,\alpha_i.$$

Then

$$E\left[\frac{y_{it}}{\lambda_{it}(\theta)} - \frac{y_{it-1}}{\lambda_{it-1}(\theta)}\middle| x_{i1},\ldots,x_{it-1}\right] = 0$$

A set of sequential moment conditions.

Construct GMM estimator from sample counterparts to unconditional moments.

Semiparametric binary choice

Fix $T = 2$.

Suppose that

$$P[y_{it} = 1 | x_i, \alpha_i] = F_i(\alpha_i + x'_{it}\theta_0)$$

for F_i unknown but strictly increasing.

Then

$$\begin{aligned} P[y_{i1} = 1|x_i, \alpha_i] > P[y_{i2} = 1|x_i, \alpha_i] &\Longrightarrow \alpha_i + x'_{i1}\theta_0 > \alpha_i + x'_{i2}\theta_0 \\ P[y_{i1} = 1|x_i, \alpha_i] < P[y_{i2} = 1|x_i, \alpha_i] &\Longrightarrow \alpha_i + x'_{i1}\theta_0 < \alpha_i + x'_{i2}\theta_0. \end{aligned}$$

That is,

$$\begin{aligned} E[\Delta y_i|x_i] > 0 &\Longrightarrow \Delta x'_i\theta > 0 \\ E[\Delta y_i|x_i] < 0 &\Longrightarrow \Delta x'_i\theta < 0, \end{aligned}$$

or simply

$$\text{sign}\left\{E\big[\Delta y_i|x_i\big]\right\} = \text{sign}\left\{\Delta x'_i\theta_0\right\},$$

which is a ‘moment condition’ that does not involve fixed effects.

The scale of θ is undetermined here; so we normalize $\|\theta\| = 1$.

Maximum score (Manski 1987)

The corresponding estimator maximizes

$$\sum_i \Delta y_i \ \text{sign}\{\Delta x_i'\theta\} = \sum_i \Delta y_i \ \text{sign}\{\Delta x_i'\theta\}\, 1(\Delta y_i \neq 0)$$

over the unit sphere Θ. This is a conditional maximum-score estimator; it uses only subsample of movers, like conditional logit.

The large-sample version is $Q(\theta) = E\left[\Delta y_i \,\text{sign}\{\Delta x_i'\theta\}\right]$. With

$$\mathcal{X}(\theta') = \{z \in \mathcal{R}^{\dim\theta} : \text{sign}(z'\theta') \neq \text{sign}(z'\theta)\}$$

we have

$$\begin{aligned} Q(\theta) - Q(\theta') &= E\{E[\Delta y_i|\Delta x_i]\left(\text{sign}\{\Delta x_i\theta\}] - \text{sign}\{\Delta x_i\theta'\}]\right)\} \\ &= 2\int_{\mathcal{X}(\theta')} E[\Delta y_i|\Delta x_i]\,\text{sign}\{\Delta x_i\theta\}]\,dF(\Delta x_i) \\ &= 2\int_{\mathcal{X}(\theta')} |E[\Delta y_i|\Delta x_i]|\,dF(\Delta x_i) \geq 0. \end{aligned}$$

For point identification (up to scale) we therefore need that $P(\Delta x_i \in \mathcal{X}(\theta')) > 0$ for all $\theta' \in \Theta$, which is implied by a large-support condition.

Smoothed maximum score (Horowitz 1992)

Maximum-score objective function is non-smooth.

$\|\widehat{\theta} - \theta\| = O_p(N^{-1/3})$ and the limit distribution is non-normal.

Bootstrap does not work.

The smoothed maximum-score estimator maximizes

$$\sum_i \frac{1 + \Delta y_i}{2} \; K\left(\frac{\Delta x_i' \theta}{h}\right)$$

where K is chosen so that $\lim_{a \to \infty} K(a) = 1$ and $\lim_{a \to \infty} K(-a) = 0$ (a CDF, for example).

This objective function is smooth if K is smooth.

Under sufficient smoothness on the distributions of $\varepsilon_{it} | x_i$ and of x_i, this estimator can converge faster (almost as fast as a nonparametric density estimator; $N^{-2/5}$) and has a normal limit distribution.

Censored-regression

A linear model with censoring:

$$y_{it} = \max(\alpha_i + x_{it}'\theta + \varepsilon_{it}, 0).$$

Assume strict stationarity of ε_{it} and $E[\varepsilon_{it}|x_i, \alpha_i] = 0$.

Truncation makes the least-squares inconsistent because, for observations for which both y_{i1} and y_{i2} are uncensored,

$$\varepsilon_{i1} \sim F(\varepsilon_{i1}|\varepsilon_{i1} \geq -\alpha_i - x_{i1}'\theta) \quad \text{while} \quad \varepsilon_{i2} \sim F(\varepsilon_{i2}|\varepsilon_{i2} \geq -\alpha_i - x_{i2}'\theta).$$

Therefore,

$$E[\Delta y_i - \Delta x_i'\theta | x_i, y_{i1} > 0, y_{i2} > 0] = E[\Delta\varepsilon_i | x_i, \varepsilon_{i1} \geq -\alpha_i - x_{i1}'\theta, \varepsilon_{i2} \geq -\alpha_i - x_{i2}'\theta],$$

which is non-zero, in general.

Pairwise differencing (Honoré 1992)

Least-squares can be fixed because the mean-zero condition can be restored.

The idea is to equalize the censoring threshold across periods.

Consider the uncensored subsample. Note that

$$E[\Delta\varepsilon_i|x_i,\alpha_i,\varepsilon_{i1}\geq\max\{-\alpha_i-x'_{i1}\theta,-\alpha_i-x'_{i2}\theta\},\varepsilon_{i2}\geq\max\{-\alpha_i-x'_{i1}\theta,-\alpha_i-x'_{i2}\theta\}]$$

is zero by stationarity.

The errors are not observed, but

$$\begin{aligned}\varepsilon_{i1} > -\alpha_i - x'_{i1}\theta &\iff y_{i1} > 0\\ \varepsilon_{i1} > -\alpha_i - x'_{i2}\theta &\iff y_{i1} > (x_{i1}-x_{i2})'\theta\end{aligned}$$

and, similarly,

$$\begin{aligned}\varepsilon_{i2} > -\alpha_i - x'_{i1}\theta &\iff y_{i2} > (x_{i2}-x_{i1})'\theta\\ \varepsilon_{i2} > -\alpha_i - x'_{i2}\theta &\iff y_{i2} > 0\end{aligned}$$

Estimation

Can build an estimator from the moment condition

$$E[\Delta y_i - \Delta x_i'\theta | x_i, y_{i1} > \max\{0, -\Delta x_i'\theta\}, y_{i2} > \max\{0, \Delta x_i'\theta\}] = 0$$

Trimmed least squares:

$$\sum_i \left(\Delta y_i - \Delta x_i'\theta\right)^2 \tau_i(\theta)$$

Trimmed least absolute deviations:

$$\sum_i \left|\Delta y_i - \Delta x_i'\theta\right| \tau_i(\theta)$$

For trimming function

$$\tau_i(\theta) = 1\left[y_{i1} \geq \max\left(0, -\Delta x_i'\theta\right), y_{i2} \geq \max\left(0, \Delta x_i'\theta\right)\right].$$

Large-T panels

Limitations under classical asymptotics

Our overview of the literature on classical asymptotics shows that the incidental-parameter problem is alove and well.

The proposed solutions are highly model specific

Many parameters of interest are simply not point identified under this paradigm.

Yet, there are panels where $T < N$, but $T >> 2$.

Think about the PSID, for example.

Can think of incidental-parameter problem as a small-T bias.

Consider corrections for it that are justified under rectangular-array asymptotics.

Rectangular-array asymptotics have $N/T \to c$ for a finite constant c.

Intuition and bias expansion as $T \to \infty$

Typically, the bias in MLE is $O(T^{-1})$.

In smooth problems,

$$\text{plim}_{N\to\infty}\widehat{\theta} - \theta_0 = \frac{B}{T} + o(T^{-1}),$$

The variance is $O((NT)^{-1})$, and so

$$\sqrt{NT}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(c\,B, I^{-1}),$$

eventhough we have consistency, as $N, T \to \infty$ with $N/T \to c^2$, and confidence intervals are useless.

Bias correction as $T \to \infty$

If we can we can estimate B by $\widehat{B}$. Then

$$\widetilde{\theta} = \widehat{\theta} - \frac{\widehat{B}}{T} \xrightarrow{p} \theta_0 + o(T^{-1});$$

so the bias is of reduced order.

Furthermore,

$$\sqrt{NT}(\widetilde{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I^{-1})$$

Confidence intervals are correctly centered if N, T grow at the same rate.

The correction shifts back the asymptotic distribution but leaves its variance unaltered.

In general, B is a complicated function of unknown parameters.

Hahn and Newey (2004), Hahn and Kuersteiner (2011).

Many normal means

With $y_{it} \sim \mathcal{N}(\alpha_i, \theta)$, the MLE is

$$\widehat{\theta} \xrightarrow{p} \theta - \frac{\theta}{T}, \qquad \text{as } N \to \infty;$$

so $B = -\theta$ and the higher-order bias is zero (expansion is exact).

A plug-in based bias-corrected estimator is

$$\widehat{\theta} + \frac{\widehat{\theta}}{T} \xrightarrow{p} \left(\theta_0 - \frac{\theta}{T}\right) + \frac{1}{T}\left(\theta - \frac{\theta}{T}\right) = \theta - \frac{\theta}{T^2}.$$

The bias is of smaller order but is not zero:

The plug-in estimator of B has itself a bias that is $O(T^{-1})$.

The asymptotic variance is $2\theta^2$, so a plug-in estimator of this variance also has bias

$$2\widehat{\theta}^2 \xrightarrow{p} 2\left(\theta - \frac{\theta}{T}\right)^2 = 2\theta^2 + O(T^{-1})$$

$$2\widetilde{\theta}^2 \xrightarrow{p} 2\left(\theta - \frac{\theta}{T^2}\right)^2 = 2\theta^2 + O(T^{-2}).$$

(in more general models, the latter will not improve on the former.)

Delete-one jackknife correction (Hahn and Newey 2004)

Data are i.i.d. across time given α_i.

Let $\widehat{\theta}_{-t}$ be the MLE using periods $\{1, 2, \dots, t-1, t+1, \dots, T\}$, i.e., deleting cross-section t. Then

$$\widehat{\theta}_{-t} \xrightarrow{p} \theta - \frac{\theta}{T-1}.$$

Hence,

$$(T-1)(\widehat{\theta}_{-t} - \widehat{\theta}) \xrightarrow{p} (T-1)\left(\theta - \frac{\theta}{T-1} - \theta + \frac{\theta}{T}\right) = -\frac{\theta}{T}$$

for all t. Averaging across all possible subpanels equally gives

$$\frac{T-1}{T}\sum_t (\widehat{\theta}_{-t} - \widehat{\theta}) \xrightarrow{p} -\frac{\theta_0}{T}.$$

The delete-one (panel) jackknife is

$$\widetilde{\theta} = \widehat{\theta} - \frac{T-1}{T}\sum_t (\widehat{\theta}_{-t} - \widehat{\theta}) = T\widehat{\theta} - (T-1)\,\overline{\theta} \xrightarrow{p} \theta,$$

for $\overline{\theta} = T^{-1}\sum_{t=1}^{T} \widehat{\theta}_{-t}$.

The bias is completely elimitated in this example.

Split-panel jackknife (Dhaene and Jochmans 2015)

When data are not i.i.d. across time (as would be expected with panel data), the delete-one jackknife does not work.

Estimate θ from two (non-overlapping) subpanels of adjacent observations that partition $\{1, 2, \ldots, T\}$.

Suppose that T is even. Then we take subpanels

$$\{1, 2, \ldots, T/2\} \quad \text{and} \quad \{T/2 + 1, \ldots, T\},$$

and get estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$.

Under stationarity,

$$\text{plim}_{N\to\infty} \widehat{\theta}_1 - \theta = \frac{B}{T/2} + o(T^{-1}), \qquad \text{plim}_{N\to\infty} \widehat{\theta}_2 - \theta = \frac{B}{T/2} + o(T^{-1}),$$

and so

$$\text{plim}_{N\to\infty} 2\widehat{\theta} - \frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2} = \theta + o(T^{-1}).$$

Linear autoregression

An expansion of the Nickell bias gives

$$\widehat{\theta} - \theta = -\frac{1+\theta}{T} + O\left(\frac{1}{T^2}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right).$$

				bias					confidence		
N	T	$\widehat{\theta}$	$\widetilde{\theta}_{\mathrm{HK}}$	$\widetilde{\theta}_{1/2}$	$\dot{\theta}_{1/2}$	$\widehat{\theta}_{\mathrm{AB}}$	$\widehat{\theta}$	$\widetilde{\theta}_{\mathrm{HK}}$	$\widetilde{\theta}_{1/2}$	$\dot{\theta}_{1/2}$	$\widehat{\theta}_{\mathrm{AB}}$
100	4	−.413	−.141	−.076	−.176	−.054	.000	.495	.682	.273	.923
100	6	−.278	−.074	−.019	−.097	−.047	.000	.702	.815	.509	.910
100	8	−.206	−.044	.001	−.058	−.039	.000	.815	.848	.702	.910
100	12	−.134	−.021	.008	−.027	−.031	.001	.897	.866	.853	.900
20	20	−.081	−.010	.005	−.012	−.089	.595	.947	.903	.935	.613
50	50	−.031	−.002	.001	−.002	−.033	.592	.950	.934	.939	.603
100	100	−.015	.000	.000	.000	−.016	.596	.948	.939	.941	.605

Note that GMM breaks down under double asymptotics.

Autoregressive probit

$$y_{it} = 1(\alpha_{i0} + \theta_0 y_{it-1} + \varepsilon_{it} > 0), \qquad \varepsilon_{it} \sim \mathcal{N}(0,1).$$

Data generated with $N = 100$, $\theta_0 = .5$, $\alpha_{i0} \sim \mathcal{N}(0,1)$.

stationary y_{i0}.

	bias			confidence			validity	
T	$\widehat{\theta}$	$\widetilde{\theta}_{1/2}$	$\dot{\theta}_{1/2}$	$\widehat{\theta}$	$\widetilde{\theta}_{1/2}$	$\dot{\theta}_{1/2}$	$\tilde{t}_{1/2}$	$\dot{t}_{1/2}$
6	$-.618$	.248	$-.272$	.031	.833	.895	.959	.929
8	$-.456$	.078	$-.162$	.079	.917	.889	.956	.951
12	$-.300$	.021	$-.074$	.194	.934	.923	.962	.962
18	$-.197$	.008	$-.031$	.354	.943	.943	.954	.954

A validity test based on the jackknife

Still consider T to be even for simplicity; so subpanels are both equally long.

Let

$$\widehat{r} \equiv (\widehat{\theta}_{S_1} - \widehat{\theta}) - (\widehat{\theta}_{S_2} - \widehat{\theta}).$$

If the expansion holds for the same B in both subpanels, then

$$\tilde{t} \equiv \frac{NT}{4} \, \widehat{r}' \, \widehat{I} \, \widehat{r} \xrightarrow{d} \chi^2_{\dim \theta},$$

which is a Wald-like statistic.

Non-stationarity

		bias			confidence		validity	
T	$\widehat{\theta}$	$\widetilde{\theta}_{1/2}$	$\dot{\theta}_{1/2}$	$\widehat{\theta}$	$\widetilde{\theta}_{1/2}$	$\dot{\theta}_{1/2}$	$\tilde{t}_{1/2}$	$\dot{t}_{1/2}$
				$y_{i0} = 0$				
6	$-.525$	.305	$-.213$	.083	.740	.936	.910	.906
8	$-.394$	.119	$-.126$	.143	.886	.928	.921	.937
12	$-.268$	.038	$-.061$	.259	.930	.944	.936	.948
18	$-.183$	.013	$-.029$	.404	.943	.945	.945	.952
				$y_{i0} = 1$				
6	$-.569$	.273	$-.242$	.054	.791	.914	.945	.921
8	$-.423$	.099	$-.142$	.112	.904	.912	.953	.952
12	$-.282$	.030	$-.066$	.233	.936	.933	.952	.954
18	$-.191$	.008	$-.032$	.375	.940	.944	.951	.953

Likelihood and score corrections

Rather than 'fixing' $\widehat{\theta}$ we can adjust the likelihood or score equation.

Remember the adjusted likelihood, where we calculated

$$b(\theta, \alpha_i) = E_{\theta,\alpha_i}\left[\frac{\partial \ell_i(\theta, \widehat{\alpha}_i(\theta))}{\partial \theta}\right] = O(1)$$

in models where $b(\theta, \alpha_i)$ did not depend on α_i.

In the general case, we can consider an adjustment of the plug-in form

$$\sum_{i=1}^{N}\left[\frac{\partial \ell_i(\theta, \widehat{\alpha}_i(\theta))}{\partial \theta} - b(\theta, \widehat{\alpha}_i(\theta))\right],$$

whose root will have a correctly-centered limit distribution under double asymptotics.

Bias approximation in more detail

Typically,

$$b(\theta, \alpha_i(\theta)) = B(\theta, \alpha_i(\theta)) + O(T^{-1}).$$

The fixed-effect version is

$$b(\theta, \widehat{\alpha}_i(\theta)) = B(\theta, \widehat{\alpha}_i(\theta)) + O(T^{-1}).$$

As $T \to \infty$,

$$\widehat{\alpha}_i(\theta) - \alpha_i(\theta) = \frac{\beta_i}{T} + \frac{1}{T}\sum_{t=1}^{T} \chi_{it} + o_p(T^{-1}), \qquad \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \chi_{it} \xrightarrow{d} \mathcal{N}(0, \Sigma_i).$$

Hence, $b(\theta, \widehat{\alpha}_i(\theta)) - b(\theta, \alpha_i(\theta))$ equals

$$\frac{\nabla_{\alpha_i} B(\theta, \alpha_i(\theta))\beta_i + \mathrm{tr}\left[\nabla_{\alpha_i \alpha_i'} B(\theta, \lambda_i(\theta))\Sigma_i\right]}{T} + O(T^{-1}) = O(T^{-1}).$$

Two-period logit

Chamberlain's two-period logit model

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

Here we know that $\text{plim}_{N\to\infty}\widehat{\theta}=2\theta$ from before.

Remember that only movers contribute to the likelihood:

$$\begin{aligned}\widehat{\alpha}_i(\theta) &= -\psi/2 \quad \text{if } y_{i1}+y_{i2}=1\\ \widehat{\alpha}_i(\theta) &= \ \pm\infty \quad \text{if } y_{i1}+y_{i2}\neq 1.\end{aligned}$$

The concentrated score is

$$\widehat{s}(\theta)=\frac{\sum_i \ell_i(\theta,\widehat{\alpha}_i(\theta))}{\partial\theta}=\frac{n_{01}}{1+e^{\theta/2}}-\frac{n_{10}}{1+e^{-\theta/2}},$$

for n_{01} and n_{10} the number of movers in an out the sample.

Bias calculation

We have

$$E_{\theta,\alpha}[n_{01}] = \sum_{i=1}^{N} P_{\theta,\alpha_i}(y_{i1}=0, y_{i2}=1|\alpha_i) = \sum_{i=1}^{N} \frac{e^{-\alpha_i}}{1+e^{-\alpha_i}} \frac{1}{1+e^{-\alpha_i-\theta}}$$

$$E_{\theta,\alpha}[n_{10}] = \sum_{i=1}^{N} P_{\theta,\alpha_i}(y_{i1}=1, y_{i2}=0|\alpha_i) = \sum_{i=1}^{N} \frac{1}{1+e^{-\alpha_i}} \frac{e^{-\alpha_i-\theta}}{1+e^{-\alpha_i-\theta}}$$

Note that

$$E_{\theta,\alpha}[n_{10}] = E_{\theta,\alpha}[n_{01}]\, e^{-\theta}.$$

So,

$$E_{\theta,\alpha}[s(\theta)] = \frac{E_{\theta,\alpha}[n_{01}]}{1+e^{\theta/2}} - \frac{E_{\theta,\alpha}[n_{10}]}{1+e^{-\theta/2}} = E_{\theta,\alpha}[n_{01}]\, \frac{1-e^{-\theta/2}}{1+e^{\theta/2}},$$

which clearly depends on the α_i.

We can adjust the score by considering

$$\dot{s}_1(\theta) = \widehat{s}(\theta) - E_{\theta,\widehat{\alpha}(\theta)}[\widehat{s}(\theta)] = \frac{n_{01} - E_{\theta,\widehat{\alpha}(\theta)}[n_{01}]}{1+e^{\theta/2}} - \frac{n_{10} - E_{\theta,\widehat{\alpha}(\theta)}[n_{10}]}{1+e^{-\theta/2}}$$

Bias iteration

We can equally calculate the bias of $\dot{s}_1(\theta)$.

Moreover, we can iterate on

$$\dot{s}_k(\theta) = \dot{s}_{k-1}(\theta) - E_{\theta,\widehat{\alpha}(\theta)}[\dot{s}_{k-1}(\theta)]$$

At iteration k we have

$$\dot{s}_k(\psi) = \widehat{s}(\psi) - \left(1 - (1-b_\psi)^k\right) \frac{a_\theta}{b_\theta}\, E_{\theta,\widehat{\alpha}(\theta)}[n_{01}]$$

for

$$a_\theta = \frac{1-e^{-\theta/2}}{1+e^{\theta/2}}, \qquad b_\theta = \frac{1+e^{-\theta}}{(1+e^{-\theta/2})^2}, \qquad E_{\theta,\widehat{\alpha}(\theta)} n_{01} = \frac{n_{01}+n_{10}}{\left(1+e^{-\theta/2}\right)^2},$$

and so

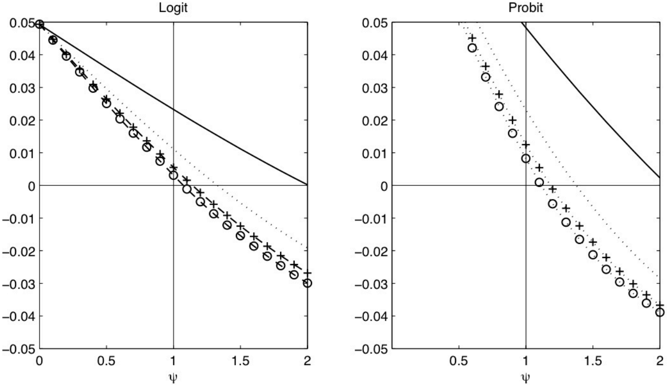
$$E_{\theta,\alpha}\dot{s}_k(\theta) = a_\theta\,(1-b_\theta)^k\,\mathbb{E}_{\theta,\alpha}[n_{01}] \xrightarrow{k\to\infty} 0,$$

as $0 < b_\theta < 1$.
In the limit,

$$\dot{s}_\infty(\psi) = \frac{n_{01}}{1+e^{\psi}} - \frac{n_{10}}{1+e^{-\psi}},$$

which equals the score of the conditional likelihood.

Score plots for $k = 0, 1, 2, 3$



Scores are for large N.

Numerical example

	mean bias				median bias			
T	$\widehat{\theta}$	$\dot{\theta}$	$\ddot{\theta}$	$\overline{\theta}$	$\widehat{\theta}$	$\dot{\theta}$	$\ddot{\theta}$	$\overline{\theta}$
2	1.2080	.4275	.2294	.1040	1.0473	.3282	.1457	.0237
4	.4349	.0813	.0266	.0177	.4091	.0667	.0172	.0013
6	.2467	.0333	.0088	.0061	.2381	.0279	.0006	.0010
12	.1080	.0086	.0041	.0033	.1030	.0052	.0025	.0004
	standard deviation				interquartile range			
m	$\widehat{\theta}$	$\dot{\theta}$	$\ddot{\theta}$	$\overline{\theta}$	$\widehat{\theta}$	$\dot{\theta}$	$\ddot{\theta}$	$\overline{\theta}$
2	.8198	.5231	.4879	.4099	.9205	.5723	.5393	.4603
4	.2687	.1870	.2005	.1754	.3532	.2435	.2638	.2353
6	.1740	.1380	.1556	.1331	.2370	.1884	.2144	.1821
12	.0964	.0869	.1019	.0853	.1324	.1226	.1406	.1189

$N = 250$, $y_{it} = 1\{\alpha_i + x_{it}\theta \geq \varepsilon_{it}\}$, $\varepsilon_{it} \sim$ unit-logistic, $\alpha_i \sim \mathcal{N}(0,1)$, $x_{it} \sim \mathcal{N}(0,1)$. $\overline{\theta}$: conditional MLE.

Average marginal effects

In a binary-choice model with

$$P[y_{it} = 1 | x_i, \alpha_i] = F(\alpha_i + x_{it}\theta),$$

the marginal effect for unit i at $x_{it} = x$

$$\frac{\partial F(\alpha_i + x\theta_0)}{\partial x} = \theta\, f(\alpha_i + x\theta).$$

Nonlinearity is attractive, but dependence on α_i means we cannot get around the estimation of α_i.

The MLE uses

$$\widehat{\theta}\, f(\widehat{\alpha}_i + x\widehat{\theta}).$$

Even the infeasible estimator

$$\theta_0\, f(\widehat{\alpha}_i + x\theta)$$

will not have 'good' properties, because of the presence of $\widehat{\alpha}_i$.

Bias expansions

Estimand:

$$\mu = \lim_{N \to \infty} \frac{1}{N} \sum_i E[\mu_{it}(\theta, \alpha_i)].$$

MLE:

$$\widehat{\mu} = \frac{1}{NT} \sum_i \sum_t \mu_{it}(\widehat{\theta}, \widehat{\alpha}_i).$$

Bias in MLE comes from

- bias in $\widehat{\theta}$,
- bias in $\widehat{\alpha}_i$.

Each of these contributes a $O(T^{-1})$ term, as in

$$\text{plim}_{N \to \infty} \widehat{\mu} - \mu_0 = \frac{B_1 + B_2}{T} + O(T^{-2}),$$

say.

A delete-one jackknife correction is

$$\widetilde{\mu} = T\widehat{\mu} - \frac{T-1}{T} \sum_t \widehat{\mu}_{-t}(\widehat{\theta}_{-t}, \widehat{\alpha}_{i(-t)})$$

Asymptotic distribution

Think about α_i as draws from a distribution G.

The convergence rate of $\widehat{\mu}$ is $N^{-1/2}$, and the bias is asymptotically negligible under rectangular-array asymptotics.

Let

$$\mu_* = \frac{1}{NT} \sum_i \sum_t \mu_{it}(\theta, \alpha_i).$$

Then

$$\begin{aligned}
\mu_* &= \frac{1}{N} \sum_i E[\mu_{it}(\theta, \alpha_i)] + \frac{1}{N} \sum_i \left(\frac{1}{T} \sum_t \mu_{it}(\theta, \alpha_i) - E[\mu_{it}(\theta, \alpha_i)] \right) \\
&= \mu_0 + O_p(N^{-1/2}) + O_p((NT^{-1/2})),
\end{aligned}$$

so

$$\sqrt{N}(\mu_* - \mu) = O_p(1) + O_p(T^{-1/2})$$

Any feasible estimator will converge no faster than this:

$$\sqrt{N}(\widehat{\mu} - \mu) = O_p(1) + O_p(T^{-1/2}) + O_p(T^{-1}).$$

Inference

Bias correction tends to improve finite-sample performance.

Variance estimator based on limit result does not perform well.

Can consider adding a T^{-1} penalty to the variance estimator to account for all the sampling noise.

One option is to use an expansion of the estimating equation to account for variability in estimates of θ and α_i.

A simple numerical example is the average derivative of the survival function at zero in a Gaussian autoregression:

$$\iint \frac{\rho}{\sigma}\phi\left(\frac{\alpha+\rho x}{\sigma}\right) F_{x|\alpha}(dx)\, G(d\alpha)$$

Numerical example

		bias				sd			
N	T	$\widehat{\mu}$	$\widetilde{\mu}_{1/2}$	μ_*	$\widehat{\mu}(\theta_0)$	$\widehat{\mu}$	$\widetilde{\mu}_{1/2}$	μ_*	$\widehat{\mu}(\theta_0)$
100	4	−.096	−.023	.000	.015	.016	.027	.007	.007
100	8	−.045	−.005	.000	.006	.008	.013	.007	.006
100	12	−.028	−.002	.000	.004	.007	.010	.007	.006
100	16	−.021	−.001	.000	.003	.007	.009	.006	.006
100	24	−.013	−.001	.000	.002	.006	.008	.006	.006
50	50	−.006	−.001	.000	.001	.009	.010	.009	.009
100	100	−.003	.000	.000	.001	.006	.006	.006	.006
250	250	−.001	.000	.000	.000	.004	.004	.004	.004

		se/sd				confidence			
N	T	$\widehat{\mu}$	$\widetilde{\mu}_{1/2}$	μ_*	$\widehat{\mu}(\theta_0)$	$\widehat{\mu}$	$\widetilde{\mu}_{1/2}$	μ_*	$\widehat{\mu}(\theta_0)$
100	4	.057	.142	.994	.990	.000	.179	.946	.358
100	8	.338	.420	1.000	1.002	.000	.551	.946	.831
100	12	.571	.596	.991	.988	.005	.736	.946	.901
100	16	.702	.704	.997	1.000	.052	.818	.945	.924
100	24	.819	.811	.990	.991	.308	.880	.946	.935
50	50	.916	.913	.987	.989	.853	.920	.942	.941
100	100	.965	.963	1.000	1.000	.910	.937	.948	.946
250	250	.988	.987	1.001	1.001	.932	.946	.949	.949

		se/sd with correction				confidence with correction			
N	T	$\widehat{\mu}$	$\widetilde{\mu}_{1/2}$	μ_*	$\widehat{\mu}(\theta_0)$	$\widehat{\mu}$	$\widetilde{\mu}_{1/2}$	μ_*	$\widehat{\mu}(\theta_0)$
100	4	1.003	.944	1.015	1.112	.000	.825	.952	.453
100	8	1.000	.861	1.016	1.065	.001	.868	.949	.862
100	12	.983	.892	1.003	1.033	.030	.900	.948	.916
100	16	.992	.926	1.007	1.035	.136	.921	.947	.933
100	24	.993	.956	.997	1.016	.430	.933	.948	.941
50	50	.990	.981	.991	1.002	.878	.942	.943	.944
100	100	1.000	.996	1.002	1.006	.919	.946	.948	.948
250	250	1.001	1.001	1.002	1.004	.935	.948	.949	.950

Two-way models

Two-way models

So far, only fixed-effect heterogeneity in one dimension. Panel data has two dimensions.

One motivation: time effects

Another motivation: linked data

-employer/employee data

-importer/exporter trade data.

The linear model with two-way effects is just

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

One binary-choice formulation would be

$$P(y_{it} = 1 | x_i, \alpha_i, \gamma_t) = F(x_{it}'\theta + \alpha_i + \gamma_j)$$

Another would be $P(y_{it} = 1 | x_i, \alpha_i, \gamma_t) = F(x_{it}'\theta)\, \alpha_i\, \gamma_j$.

Linear model

Simple model

$$y_{it} = \alpha_i + \gamma_j + x_{it}'\theta + \varepsilon_{it}.$$

Averaging in each dimension gives

$$\begin{aligned}
\overline{y}_i &= \alpha_i + \overline{\gamma} + \overline{x}_i'\theta + \overline{\varepsilon}_i \\
\overline{y}_t &= \overline{\alpha} + \gamma_t + \overline{x}_t'\theta + \overline{\varepsilon}_t,
\end{aligned}$$

and averaging in both dimension gives $\overline{y} = \overline{\alpha} + \overline{\gamma} + \overline{x}'\theta + \overline{\varepsilon}$.

Hence,

$$\begin{aligned}
y_{it} - \overline{y} &= (\alpha_i - \overline{\alpha}) + (\gamma_t - \overline{\gamma}) + (x_{it} - \overline{x})'\theta + (\varepsilon_{it} - \overline{\varepsilon}) \\
\overline{y}_i - \overline{y} &= (\alpha_i - \overline{\alpha}) + \qquad\qquad\quad + (\overline{x}_i - \overline{x})'\theta + (\overline{\varepsilon}_i - \overline{\varepsilon}) \\
\overline{y}_t - \overline{y} &= \qquad\qquad\quad +(\gamma_t - \overline{\gamma}) + (\overline{x}_t - \overline{x})'\theta + (\overline{\varepsilon}_t - \overline{\varepsilon})
\end{aligned}$$

and so, on writing $\tilde{y}_{it} = (y_{it} - \overline{y}) - (\overline{y}_i - \overline{y}) - (\overline{y}_t - \overline{y})$ etc, we have

$$\tilde{y}_{it} = \tilde{x}_{it}'\theta + \tilde{\varepsilon}_{it}$$

Double-differenced OLS is consistent under strict exogeneity.

Conditional logit

$$P(y_{it} = 1 | x_{it}, \alpha_i, \gamma_t) = F(x_{it}'\theta + \alpha_i + \gamma_j), \qquad F(a) = \frac{1}{1+e^{-a}},$$

with $y_{it} | x, \alpha, \gamma$ independent across i, t.

Consider a quad of data

$$\begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix}$$

(the two-way analog to a two-period one-way model).

A sufficient statistic exists and equals

$$y_{11} + y_{12} = 1, \qquad y_{21} + y_{22} = 1, \qquad y_{11} + y_{21} = 1, \qquad y_{12} + y_{22} = 1.$$

(note that 1 of these conditions is redundant).

This generalizes the concept of movers to the two-way model:

$$\begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix} \in \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}.$$

is the relevant subpopulation.

Conditional probabilities

Let

$$z = \frac{(y_{11} - y_{12}) - (y_{21} - y_{22})}{2}.$$

The relevant subpopulation has $z \in \{1, -1\}$.

$P(z = 1|x)$ equals

$$\frac{1}{1 + e^{-\alpha_1 - \gamma_1 - x'_{11}\theta}} \frac{e^{-\alpha_1 - \gamma_2 - x'_{12}\theta}}{1 + e^{-\alpha_1 - \gamma_2 - x'_{12}\theta}} \frac{e^{-\alpha_2 - \gamma_1 - x'_{21}\theta}}{1 + e^{-\alpha_2 - \gamma_1 - x'_{21}\theta}} \frac{1}{1 + e^{-\alpha_2 - \gamma_2 - x'_{22}\theta}}.$$

$P(z = -1|x)$ equals

$$\frac{e^{-\alpha_1 - \gamma_1 - x'_{11}\theta}}{1 + e^{-\alpha_1 - \gamma_1 - x'_{11}\theta}} \frac{1}{1 + e^{-\alpha_1 - \gamma_2 - x'_{12}\theta}} \frac{1}{1 + e^{-\alpha_2 - \gamma_1 - x'_{21}\theta}} \frac{e^{-\alpha_2 - \gamma_2 - x'_{22}\theta}}{1 + e^{-\alpha_2 - \gamma_2 - x'_{22}\theta}}.$$

Conditional probabilities

Then

$$\frac{P(z=1|x)}{P(z=1|x)+P(z=-1|x)} = \frac{1}{1+\frac{P(z=-1|x)}{P(z=-1|x)}} = \frac{1}{1+\frac{e^{-\alpha_1-\gamma_1-x_{11}'\theta}\,e^{-\alpha_2-\gamma_2-x_{22}'\theta}}{e^{-\alpha_1-\gamma_2-x_{12}'\theta}\,e^{-\alpha_2-\gamma_1-x_{21}'\theta}}}$$

$$= \frac{1}{1+e^{(x_{12}-x_{11}+x_{21}-x_{22})'\theta}}$$

$$= F\left(r'\theta\right)$$

and

$$\frac{P(z=-1|x)}{P(z=1|x)+P(z=-1|x)} = 1 - F\left(r'\theta\right)$$

for $r = (x_{11}-x_{12})-(x_{21}-x_{22})$.

This is like difference-in-differences.

Conditional likelihood

The data consists of

$$\rho = \binom{N}{2}\binom{T}{2} = \frac{N(N-1)\,T(T-1)}{4}$$

distinct quads.

For each $i = 1, 2, \ldots, \rho$, construct z_i and r_i.

We can then estimate θ by maximizing

$$\sum_{i=1}^{\rho} 1\{z_i = 1\}\,\log F(r_i'\theta) + 1\{z_i = -1\}\,\log(1 - F(r_i'\theta)).$$

Standard conditional-likelihood theory does not apply here (the default standard errors are not valid).

Asymptotics are in Jochmans (2015).

Multiplicative models

Conditional-mean restriction as

$$E[y_{it}|x_{it},\alpha_i,\gamma_t] = \varphi(x_{it};\theta)\,\alpha_i\,\gamma_t.$$

and $y_{it}|x,\alpha,\gamma$ independent across i,t.

Let

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

so

$$E\left[\tau_{it}(\theta)|x_{it},\alpha_i,\gamma_t\right] = \alpha_i\gamma_t.$$

Then

$$\begin{aligned} E\left[\tau_{11}(\theta)\tau_{22}(\theta)|x,\alpha,\gamma\right] &= \alpha_1\alpha_2\gamma_1\gamma_2 \\ E\left[\tau_{12}(\theta)\tau_{21}(\theta)|x,\alpha,\gamma\right] &= \alpha_1\alpha_2\gamma_1\gamma_2 \end{aligned}$$

and

$$E\left[\tau_{11}(\theta)\tau_{22}(\theta) - \tau_{12}(\theta)\tau_{21}(\theta)|x\right] = 0.$$

GMM

Unconditional moments follow; for example

$$q(\theta) = \sum_{i=1}^{N} \sum_{i<i'} \sum_{t=1}^{T} \sum_{t<t'} [(x_{it} - x_{i't'}) - (x_{it'} - x_{i't})] \, [\tau_{it}(\theta)\tau_{i't'}(\theta) - \tau_{it'}(\theta)\tau_{i't}(\theta)],$$

The associated GMM estimator minimizes a quadratic form like

$$q(\theta)' W q(\theta).$$

Similar asymptotics as above.

Good properties as $N, T \to \infty$ jointly at any rate.

Bias reduction

Incidental-parameter bias in MLE shows up in both dimensions.

Now (under regularity),

$$\widehat{\theta} - \theta = \frac{B}{T} + \frac{C}{N} + o(\max\{T^{-1}, N^{-1}\}) + O_p((NT)^{-1/2})$$

as $N, T \to \infty$ so that $N/T \to c^2$.

Note that this is the only asymptotic scheme that gives a well-defined limit distribution.

Here,

$$\sqrt{NT}(\widehat{\theta} - \theta) = \sqrt{\frac{N}{T}}\, B + \sqrt{\frac{T}{N}}\, C + O_p(1)$$

and we may consider correcting for both sources of bias (B and C).

Fernández-Val and Weidner (2015).

Modified likelihood

An alternative is to bias-corrected the profile likelihood.

Collect α and γ in $N+T$ vector λ.

The profile likelihood is

$$\hat{\ell}(\theta) = \ell(\theta, \hat{\lambda}(\theta))$$

and is a plug-in estimator of the target likelihood

$$\ell(\theta) = \ell(\theta, \lambda(\theta)).$$

Under sufficient regularity,

$$\hat{\lambda}(\theta) - \lambda(\theta) \approx H(\theta)^{-1} V(\theta)$$

for

$$H(\theta, \lambda) = -E\left(\frac{\partial^2 \ell(\theta, \lambda)}{\partial \lambda \partial \lambda'}\right), \qquad V(\theta, \lambda) = \frac{\partial \ell(\theta)}{\partial \lambda},$$

and $H(\theta) = H(\theta, \lambda(\theta))$ and $V(\theta) = V(\theta, \lambda(\theta))$.

Modified likelihood (cont'd)

An expansion gives

$$\beta(\theta) = E\left(\hat{\ell}(\theta) - \ell(\theta)\right) = -\frac{1}{2}\mathrm{trace}\left(H(\theta)^{-1}\Omega(\theta)\right)$$

for

$$\Omega(\theta) = E\left(V(\theta)V(\theta)'\right).$$

In general,

$$\beta(\theta) = O(N) + O(T).$$

So,

$$\tilde{\ell}(\theta) = \hat{\ell}(\theta) - \hat{\beta}(\theta)$$

will give a bias-corrected point estimator.

See Jochmans and Otsu (2016).