NONLINEAR MODELS FOR PANEL DATA

(MRes in Economics)

Koen Jochmans

Toulouse School of Economics

Last revised on March 25, 2022

Outline

1 **Short panels**
- Examples
- Recentering the score equation
- Conditional likelihood
- Ad hoc moment restrictions
- Bounds

2 **Rectangular-array asymptotics**
- Bias correction
- Parametric bootstrap
- Computation
- Average effects

Examples

Setup

Double-indexed data z_{it}.

Units $i = 1, \ldots, N$ and time periods $t = 1, \ldots, T$.

Consider parametric models:

$$f(z_{it}|\theta, \eta_i)$$

(this covers conditional distributions; set $z_{it} = (y_{it}, x_{it})$ and condition on x_{it}).

Here, θ is a common parameter while η_i is a unit-specific parameter.

Both these parameters may be vectors.

Interest is in microsettings.

Consider asymptotics where T is fixed and where T grows with N.

An example: Linear regression

Classical linear regression with unit-specific intercepts:

$$y_{it} = \eta_i + x_{it}\beta + \varepsilon_{it}, \qquad \varepsilon_{it}|x_{i1}, \ldots, x_{iT} \sim N(0, \sigma^2).$$

Here, $\theta = (\beta, \sigma^2)$ and the regressor is taken to be a scalar for simplicity.

Usual motivation from an error-component formulation.

Here, the η_i are treated as fixed. Expectations are conditional on them.

The log-likelihood function is

$$-\frac{NT}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{(y_{it} - \eta_i - x_{it}\beta)^2}{\sigma^2}.$$

Main interest lies in β.

The score equation for η_i is

$$\frac{\sum_{t=1}^{T}(y_{it} - \eta_i - x_{it}\beta)}{\sigma^2} = 0$$

and those for β and σ^2 are

$$\frac{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\,(y_{it} - \eta_i - x_{it}\beta)}{\sigma^2} = 0$$

$$\frac{1}{2\sigma^2}\left(-NT + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \eta_i - x_{it}\beta)^2}{\sigma^2}\right) = 0$$

respectively.

All observations contribute to the scores for the common parameters.

Only observations on unit i contribute to the score for η_i.

For given values of $\theta = (\beta, \sigma^2)$,

$$\hat{\eta}_i(\theta) = \frac{1}{T}\sum_{t=1}^{T}(y_{it} - x_{it}\beta) = \bar{y}_i - \bar{x}_i\beta$$

is the maximum-likelihood estimator of η_i.

Plug this into the scores for the common parameters to get

$$\frac{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\left((y_{it} - \bar{y}_i) - (x_{it} - \bar{x}_i)\beta\right)}{\sigma^2} = 0$$

$$\frac{1}{2\sigma^2}\left(-NT + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}((y_{it} - \bar{y}_i) - (x_{it} - \bar{x}_i)\beta)^2}{\sigma^2}\right) = 0$$

as profiled estimating equations for β and σ^2.

Recall from basic results on the linear model that the scores for β and σ^2 are information orthogonal.

The expectation of the score for β (at true values) is

$$\begin{aligned}
\mathbb{E}\left(\frac{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\left(\varepsilon_{it}-\bar{\varepsilon}_i\right)}{\sigma^2}\right) &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}(x_{it}\varepsilon_{it})}{\sigma^2} \\
&- \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}(x_{it}\bar{\varepsilon}_i\ )}{\sigma^2} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}(x_{it}\mathbb{E}(\varepsilon_{it}|x_{i1},\ldots,x_{iT}))}{\sigma^2} \\
&- \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}(\,\bar{x}_i\ \mathbb{E}(\varepsilon_{it}|x_{i1},\ldots,x_{iT}))}{\sigma^2} \\
&= \ 0
\end{aligned}$$

because

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

for all time periods. Regressors are strictly exogenous.

The score for β is unbiased and (in light of the information orthogonality mentioned above) the estimator

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

is unbiased and consistent for β whether T is fixed or not.

The expectation of the score for σ^2 (at true values) is

$$\frac{1}{2\sigma^2}\mathbb{E}\left(-NT + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(\varepsilon_{it} - \bar{\varepsilon}_i)^2}{\sigma^2}\right).$$

Working out the square and re-arranging yields

$$\frac{1}{2\sigma^2}\left(-NT + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(\sigma^2 - 2\sigma^2/T + \sigma^2/T)}{\sigma^2}\right) = -\frac{N}{2\sigma^2} \neq 0.$$

The score for σ^2 is biased.

The estimator

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

is inconsistent if T is treated as fixed. Its probability limit is

$$\text{plim}_{N\to\infty}\hat{\sigma}^2 = \sigma^2 - \frac{\sigma^2}{T}.$$

This is the usual degrees-of-freedom issue but, here, it persists as N grows.

Fixed effects cannot be estimated consistently for fixed T.

The score equation for β is unbiased because it is linear in $\hat{\eta}$ and the latter is unbiased for η_i.

The score equation for σ^2 is biased because it is nonlinear in $\hat{\eta}_i$ and the $\hat{\eta}_i$ have positive variance.

In large N fixed T asymptotics this phenomenon causes inconsistency of the maximum-likleihood estimator and is known as the incidental-parameter problem.

Note that information orthogonality prevents bias in the equation for σ^2 to affect estimation of β. This would not otherwise be the case.

In an autoregressive specification, where $x_{it} = y_{it-1}$ the $\hat{\eta}_i$ are biased and β is subject to the incidental-parameter problem. This particular example is known as the Nickell bias.

If (strictly-exogenous) regressors are added their slopes are also subject to the same problem because the scores and the one of the autoregressive parameter are not information orthogonal.

The above conclusion extends to nonlinear functions of the fixed effects.

For example, survival probabilities such as

$$\mathbb{P}(y_{it} > y|x_i) = \mathbb{P}(\varepsilon_{it} > y - \eta_i - x_{it}\beta|x_i) = 1 - \Phi\left(\frac{y - \eta_i - x_{it}\beta}{\sigma}\right)$$

cannot be estimated consistently for fixed T.

Comment: semiparametric version and variance estimation

The estimator for β is the usual within-group estimator:

Pooled least-squares regression of $y_{it} - \bar{y}_i$ on $x_{it} - \bar{x}_i$. This estimator only requires the conditional-mean restriction

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

to be consistent and asymptotically normal for fixed T.

Stack $y_i = (y_{i1}, \ldots, y_{iT})'$ and $x_i = (x_{i1}, \ldots, x_{iT})'$ and define the $(T-1) \times T$ matrix

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

Then within-groups is based on the moment condition

$$\mathbb{E}(x_i' M(y_i - x_i\beta)) = \mathbb{E}(x_i' M \varepsilon_i) = 0$$

for $M = D'(DD')^{-1}D$ the matrix that transforms y_i into deviations from its within-group mean.

The variance of the moment is

$$\mathbb{E}(x_i' M \varepsilon_i \varepsilon_i' M x_i)$$

and we can estimate this by

$$\frac{1}{N} \sum_{i=1}^{N} x_i' M (y_i - x_i \hat{\beta})(y_i - x_i \hat{\beta})' M x_i.$$

This is robust to heteroskedasticity and (within-group) serial correlation.

Suppose errors are uncorrelated (but may be heteroskedastic). The infeasible White-type estimator of the moment variance is

$$\frac{1}{(NT)^2} \sum_i \sum_t (x_{it} - \bar{x}_i)^2 \varepsilon_{it}^2.$$

We do not know ε_{it} so replace it by within-group residuals

$$\hat{\varepsilon}_{it} = y_{it} - \hat{\eta}_i - x_{it} \hat{\beta}.$$

Suppose we know β, so that

$$\hat{\varepsilon}_{it} = y_{it} - \hat{\eta}_i - x_{it}\beta = \varepsilon_{it} - (\hat{\eta}_i - \eta_i) = \varepsilon_{it} - \bar{\varepsilon}_i.$$

Then

$$\mathbb{E}(\hat{\varepsilon}_{it}^2 | x_i) = \mathbb{E}(\varepsilon_{it}^2 - 2\varepsilon_{it}\bar{\varepsilon}_i + \bar{\varepsilon}_i^2 | x_i) = \sigma_{it}^2 - 2\frac{\sigma_{it}^2}{T} + \frac{1}{T}\frac{\sum_{s=1}^{T}\sigma_{is}^2}{T}.$$

The bias does not vanish for fixed T.

It follows that

$$\frac{1}{(NT)^2}\sum_i\sum_t (x_{it} - \bar{x}_i)^2\hat{\varepsilon}_{it}^2 - \frac{1}{(NT)^2}\sum_i\sum_t (x_{it} - \bar{x}_i)^2\varepsilon_{it}^2 \overset{p}{\nrightarrow} 0$$

and White-type standard errors are inconsistent.

An example: Exponential regression

Poisson model has pmf

$$\frac{(\eta_i e^{x_{it}\theta})^{y_{it}}\ e^{-\eta_i e^{x_{it}\theta}}}{y_{it}!}.$$

The log-likelihood is

$$\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}\log(\eta_i e^{x_{it}\theta}) - \eta_i e^{x_{it}\theta} - \log(y_{it}!).$$

The score equation for η_i is

$$\sum_{t=1}^{T}\left(\frac{y_{it}}{\eta_i} - e^{x_{it}\theta}\right) = 0$$

so that

$$\hat{\eta}_i(\theta) = \frac{\sum_{t=1}^{T} y_{it}}{\sum_{t=1}^{T} e^{x_{it}\theta}}.$$

The (profile) score for θ is

$$\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\left(y_{it} - \hat{\eta}_i(\theta)e^{x_{it}\theta}\right).$$

Note that (at true values) $\mathbb{E}_\theta(y_{it}|x_{i1},\ldots,x_{iT}) = \eta_i e^{x_{it}\theta}$ and so

$$\mathbb{E}_\theta(\hat{\eta}_i(\theta)|x_{i1},\ldots,x_{iT}) = \eta_i.$$

Therefore,

$$\mathbb{E}_\theta\left(\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\left(y_{it} - \hat{\eta}_i(\theta)e^{x_{it}\theta}\right)\right) = 0$$

and the profile-score equation is unbiased for θ.

The log-likelihood is globally concave and the maximum-likelihood estimator is consistent and asymptotically normal (under regularity conditions) for fixed T.

The above extends to semiparametric specifications where we only presume that

$$\mathbb{E}_\theta(y_{it}|x_{i1},\ldots,x_{iT}) = \eta_i e^{x_{it}\theta}.$$

Indeed, the above implies that

$$\mathbb{E}_\theta\left(\frac{y_{it}}{e^{x_{it}\theta}}\middle| x_{i1},\ldots,x_{iT}\right) = \eta_i, \qquad \mathbb{E}_\theta\left(\frac{\sum_{t=1}^T y_{it}}{\sum_{t=1}^T e^{x_{it}\theta}}\middle| x_{i1},\ldots,x_{iT}\right) = \eta_i,$$

and so

$$\mathbb{E}_\theta\left(y_{it} - \left(\frac{\sum_{s=1}^T y_{is}}{\sum_{s=1}^T e^{x_{is}\theta}}\right)e^{x_{it}\theta}\middle| x_{i1},\ldots,x_{iT}\right) = 0.$$

The poisson profile score is one unconditional version of this conditional moment condition.

An example: Logistic regression

A simple two-period logit model has

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

Here, θ is the log-odds ratio.

The log-likelihood is

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

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 η_i is

$$(y_{i1} + y_{i2}) - (F(\eta_i) + F(\eta_i + \theta)) = 0.$$

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

and so

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

Stayers do not carry information about θ, 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 likelihood is

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

The profile-score equation is

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

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

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

and so maximum-likelihood is inconsistent.

Here we have used that

$$\text{plim}_{N\to\infty}\frac{n_{10}}{n_{01}} = \frac{\text{plim}_{N\to\infty} n_{10}}{\text{plim}_{N\to\infty} n_{01}} = \frac{\lim_{N\to\infty} {}^1\!/_N \sum_{i=1}^N \mathbb{P}(\Delta y_i = -1)}{\lim_{N\to\infty} {}^1\!/_N \sum_{i=1}^N \mathbb{P}(\Delta y_i = \quad 1)} = e^{-\theta},$$

which follows from the observation that

$$\lim_{N\to\infty} {}^1\!/_N \sum_{i=1}^N \mathbb{P}(\Delta y_i = -1) = \lim_{N\to\infty} {}^1\!/_N \sum_{i=1}^N \frac{1}{1+e^{-\eta_i}} \frac{e^{-(\eta_i+\theta)}}{1+e^{-(\eta_i+\theta)}},$$

and

$$\lim_{N\to\infty} {}^1\!/_N \sum_{i=1}^N \mathbb{P}(\Delta y_i = \quad 1) = \lim_{N\to\infty} {}^1\!/_N \sum_{i=1}^N \frac{e^{-\eta_i}}{1+e^{-\eta_i}} \frac{1}{1+e^{-(\eta_i+\theta)}},$$

so that

$$\lim_{N\to\infty} {}^1\!/_N \sum_{i=1}^N \mathbb{P}(\Delta y_i = -1) = e^{-\theta} \left(\lim_{N\to\infty} {}^1\!/_N \sum_{i=1}^N \mathbb{P}(\Delta y_i = \quad 1) \right).$$

Recentering the score equation

Recentering the score equation

The log-likelihood is

$$\sum_{i=1}^{N} \ell_i(\theta, \eta_i), \qquad \ell_i(\theta, \eta_i) = \sum_{t=1}^{T} \log f(z_{it}|\theta, \eta_i).$$

The profile log-likelihood is

$$\hat{\ell}(\theta) = \sum_{i=1}^{N} \hat{\ell}_i(\theta), \qquad \hat{\ell}_i(\theta) = \ell_i(\theta, \hat{\eta}_i(\theta)).$$

The incidental-parameter problem manifests itself as

$$\sum_{i=1}^{N} \mathbb{E}_{\theta,\eta_i} \left(\frac{\partial \ell_i(\theta, \hat{\eta}_i(\theta))}{\partial \theta} \right) = O(N).$$

(At least in the parametric setting) the bias can always be calculated for given values θ, η_i, if need be by simulation.

If the profile-score bias does not depend on η_i we can correct for it exactly.

An example: Linear model

Ignore the covariate and consider

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

Then

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

Here,

$$\mathbb{E}_{\theta,\eta_i} \left(\frac{\sum_{t=1}^{T} (y_{it} - \bar{y}_i)^2}{2\theta^2} \right) = \frac{T\theta - \theta}{2\theta^2} = \frac{T-1}{2\theta}$$

and so the bias is

$$\sum_{i=1}^{N} \mathbb{E}_{\theta,\eta_i} \left(\frac{\partial \ell_i(\theta, \hat{\eta}_i(\theta))}{\partial \theta} \right) = -\frac{N}{2\theta}.$$

This is a function of θ only.

By construction the re-centered profile score equation

$$\sum_{i=1}^{N} \frac{\partial \ell_i(\theta, \hat{\eta}_i(\theta))}{\partial \theta} + \frac{N}{2\theta} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \bar{y}_i)^2}{2\theta^2} - \frac{N(T-1)}{2\theta} = 0$$

is unbiased.

Its solution is

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

which incorporates the required degrees-of-freedom correction.

An example: Autoregression

Now consider

$$y_{it} = \eta_i + \rho y_{it-1} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim N(0, \sigma^2).$$

The profile log-likelihood (conditional on the initial observation y_{i0}) is

$$-\frac{NT}{2}\log\sigma^2 - \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left((y_{it}-\bar{y}_i) - \rho\left(y_{it-1}-\bar{y}_{i-}\right)\right)^2}{2\sigma^2},$$

where

$$\bar{y}_i = {}^1\!/\!_T \sum_{t=1}^{T} y_{it}, \qquad \bar{y}_{i-} = {}^1\!/\!_T \sum_{t=1}^{T} y_{it-1}.$$

The profile scores for $\theta = (\rho, \sigma^2)$ are

$$\frac{\partial\hat{\ell}(\rho,\sigma^2)}{\partial\rho} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it-1}-\bar{y}_{i-}\right)\left(y_{it}-\rho\left(y_{it-1}-\bar{y}_{i-}\right)\right)}{\sigma^2}$$

$$\frac{\partial\hat{\ell}(\rho,\sigma^2)}{\partial\sigma^2} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-\rho\left(y_{it-1}-\bar{y}_{i}\right)\right)^2}{2\sigma^4} - \frac{NT}{2\sigma^2}$$

We focus on the former. The latter is biased in the same way as before.

Here, the key calculation is that of

$$\mathbb{E}\left(\sum_{t=1}^{T}(y_{it-1}-\bar{y}_{i-})\,\varepsilon_{it}\right)=-\sum_{t=1}^{T}\mathbb{E}\left(\bar{y}_{i-}\,\varepsilon_{it}\right).$$

Now (ignoring the fixed effect),

$$\begin{aligned}
y_{i0} &= y_{i0} \\
y_{i1} &= \rho\, y_{i0} + \varepsilon_{i1} \\
y_{i2} &= \rho^2\, y_{i0} + \rho\,\varepsilon_{i1} + \varepsilon_{i2} \\
y_{i3} &= \rho^3\, y_{i0} + \rho^2\,\varepsilon_{i1} + \rho\,\varepsilon_{i2} + \varepsilon_{i3} \\
&\vdots \\
y_{iT-1} &= \rho^{T-1} y_{i0} + \rho^{T-2}\varepsilon_{i1} + \rho^{T-3}\varepsilon_{i2} + \rho^{T-4}\varepsilon_{i3} + \cdots + \varepsilon_{iT-1}
\end{aligned}$$

and so (ignoring the fixed effect and initial condition)

$$\bar{y}_{i-} = \frac{\varepsilon_{i1}(1+\rho+\cdots+\rho^{T-2})}{T} + \frac{\varepsilon_{i2}(1+\rho+\cdots+\rho^{T-3})}{T} + \cdots + \frac{\varepsilon_{iT-1}}{T}.$$

An accounting exercise then yields

$$-\sum_{t=1}^{T}\mathbb{E}\left(\bar{y}_{i-}\,\varepsilon_{it}\right)=-\sigma^{2}\sum_{t=1}^{T-1}\frac{T-t}{T}\rho^{t-1}.$$

The score bias is thus again independent of the fixed effects (and, surprisingly, the initial values).

Re-centered profile scores are then

$$\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it-1}-\bar{y}_{i-})\left(y_{it}-\rho\left(y_{it-1}-\bar{y}_{i}\right)\right)+\sigma^{2}\sum_{t=1}^{T-1}\left({}^{(T-t)}/_{T}\right)\rho^{t-1}}{\sigma^{2}}$$

$$\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\rho\left(y_{it-1}-\bar{y}_{i}\right))^{2}}{2\sigma^{4}}-\frac{N(T-1)}{2\sigma^{2}}$$

Other examples

Weibull regression

Gamma regression

Inverse Gaussian regression

In the general case, the bias will be a function of η_i.

This case is taken up below.

Conditional likelihood

A statistic is sufficient for η_i if, conditional on it, the likelihood no longer depends on η_i.

Then the conditional likelihood can (subject to regularity conditions) be used for estimation and inference of θ.

The score of the conditional likelihood is an unbiased estimating equation for θ.

An example: Linear model

Again,

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

Note that

$$\bar{y}_i \sim N(\eta_i, {}^{\theta}/_{T}).$$

The log-density of $y_{i1,\ldots,y_{iT}}$ given $\bar{y}_i$ is

$$-\frac{T}{2}\log\theta - \frac{\sum_{t=1}^{T}(y_{it}-\eta_i)^2}{2\theta} + \frac{1}{2}\log\frac{\theta}{T} + \frac{T\,(\bar{y}_i-\eta_i)^2}{2\theta}$$

and re-arrangement yields

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

which clearly has the consistent

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

as maximizer.

An example: Logistic regression

Reconsider

$$\mathbb{P}(y_{it} = 1|x_i) = \frac{1}{1 + e^{-(\eta_i + x_{it}\theta)}} = F(\eta_i + x_{it}\theta)$$

and consider first 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 θ.

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

First,

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

is equal to

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

Now,

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

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

and so

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

Therefore,

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

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

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

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

The conditional log-likelihood is

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

The score is

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

and is clearly unbiased.

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

An example: Autoregressive logit

Now take

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

with $P_i = \mathbb{P}(y_{i0} = 1)$ unrestricted.

Suppose we observe three transitions, so we have four-wave data. (Point identification fails for less).

We have two fixed effects and look for a sufficient statistic for each.

The following two work:

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

This gives four possible sequences for $(y_{i0}, y_{i1}, y_{i2}, y_{i3})$:

- $(0, 1, 0, 1)$ (alternating in)
- $(1, 0, 1, 0)$ (alternating out)
- $(0, 0, 1, 1)$ (grouped in)
- $(1, 1, 0, 0)$ (grouped out)

We condition on the initial observation.

First take $y_{i0} = 0$. Then there are two possible sequences conditional on the sufficient statistics: 'alternating in' and 'grouped in'.

For the first,

$$\mathbb{P}(y_{i1} = 1, y_{i2} = 0, y_{i3} = 1 | y_{i0} = 0, y_{i1} + y_{i2} = 1, y_{i0} + y_{i3} = 1)$$

is equal to

$$\frac{(1 - P_i)F(\eta_i)(1 - F(\eta_i + \theta))F(\eta_i)}{(1 - P_i)F(\eta_i)(1 - F(\eta_i + \theta))F(\eta_i) + (1 - P_i)(1 - F(\eta_i))F(\eta_i)F(\eta_i + \theta)}$$

which simplifies to

$$\frac{(1 - F(\eta_i + \theta))F(\eta_i)}{(1 - F(\eta_i + \theta))F(\eta_i) + (1 - F(\eta_i))F(\eta_i + \theta)} = \frac{1}{1 + \frac{(1-F(\eta_i))F(\eta_i+\theta)}{(1-F(\eta_i+\theta))F(\eta_i)}}.$$

Here,

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

In the same way it follows that, for $y_{i0} = 1$,

$$\mathbb{P}(y_{i1} = 0, y_{i2} = 1, y_{i3} = 0 | y_{i0} = 1, y_{i1} + y_{i2} = 1, y_{i0} + y_{i3} = 1)$$

is equal to

$$\frac{1}{1+e^{-\theta}}.$$

In conclusion, the conditional probability of being an alternating sequence is $F(\theta)$.

The conditional log-likelihood equals

$$\sum_{i=1}^{N}(\{y_{i0} < y_{i3}\}\{y_{i1} > y_{i2}\} + \{y_{i0} > y_{i3}\}\{y_{i1} < y_{i2}\}) \log F(\quad \theta)$$
$$+\sum_{i=1}^{N}(\{y_{i0} < y_{i3}\}\{y_{i1} < y_{i2}\} + \{y_{i0} > y_{i3}\}\{y_{i1} > y_{i2}\}) \log F(-\theta)$$

This result implicitely uses that the choice probabilities are time invariant.

This is no longer true if we add regressors x_{it}.

However, the sufficiency argument does continue to go through conditional on $x_{i2} = x_{i3}$.

(This puts restrictions on the joint distribution of $x_{i3} - x_{i2}$ and affects the convergence rate of the implied estimator)

Ad hoc moment restrictions

An example: Additive errors

Suppose that

$$y_{it} = \eta_i + x_{it}\theta + \varepsilon_{it}, \qquad \mathbb{E}(\varepsilon_{it}|x_{i1}, \ldots, x_{it}) = 0.$$

This specification is semiparametric and allows for feedback from regressors to future outcomes.

Valid moment conditions can be derived from the fact that

$$\mathbb{E}(y_{it} - x_{it}\theta|x_{i1}, \ldots, x_{it}) = \eta_i.$$

First-differencing yields

$$\mathbb{E}(\Delta y_{it} - \Delta x_{it}\theta|x_{i1}, \ldots, x_{it-1}) = 0$$

and implies the set of sequential moment restrictions

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

Standard results on GMM can be applied here.

An example: multiplicative errors

Key here is additive separability of the errors and fixed effects.

The linear index assumption $x_{it}\theta$ can be replaced by something of the form $\varphi(x_{it}, \theta)$, for example.

A similar differencing argument can be applied to multiplicative specifications such as

$$y_{it} = \varphi(x_{it}, \theta)\, \eta_i\, \varepsilon_{it}, \qquad \mathbb{E}(\varepsilon_{it} | x_{i1}, \ldots, x_{it}) = 1.$$

Here, we would use that

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

to subsequently difference-out the fixed effects along the time dimension of the data.

An example: Censored regression

Now take

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

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

Take a two-wave panel for simplicity.

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

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

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

equals

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

and this is non-zero, in general.

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

The moment condition can be restored by artificially censoring further.

Indeed,

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

is zero by stationarity.

The errors are unobserved but

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

and, similarly,

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

We thus consider a trimmed least-squares estimator

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

An example: Binary choice

Semiparametric binary-choice model:

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

with $\varepsilon_{it}|x_i$ stationary. Let F_i be its distribution and suppose it is strictly increasing.

Take two-wave data.

Then,

$$\mathbb{P}(y_{i1} = 1|x_i) = F_i(\eta_i + x_{i1}\theta) \lesseqgtr \mathbb{P}(y_{i2} = 1|x_i) = F_i(\eta_i + x_{i2}\theta)$$

if any only if

$$\eta_i + x_{i1}\theta \lesseqgtr \eta_i + x_{i2}\theta,$$

and, using that $\mathbb{E}(\Delta y_i|x_i) = \mathbb{P}(y_{i2} = 1|x_i) - \mathbb{P}(y_{i1} = 1|x_i)$ we have

$$\text{sign}\{\mathbb{E}(\Delta y_i|x_i)\} = \text{sign}\{\Delta x_i\theta\}.$$

This suggests an M-estimator of the form

$$\sum_{i=1}^{N} \Delta y_i \operatorname{sign}\{\Delta x_i \theta\} = \sum_{i=1}^{N} \{\Delta y_i = \quad 1\} \{\Delta x_i \theta > 0\} + \{\Delta y_i = -1\} \{\Delta x_i \theta < 0\}$$

Note that:

- Only movers contribute to this criterion function.
- The criterion is invariant to changes in the scale of θ.
- The criterion is non-smooth (even in the population).

The first two points are as in the conditional logit problem (note that there the scale issue is fixed by normalized the variance of the error). Here, we need a normalization of the form $\theta/\|\theta\|$ (which, incidentally, is not very interesting in the case where we have only one regressor).

The last point implies that the estimator does not have 'standard' asymptotic properties.

Monotone index models

The maximum-score approach is powerful. Suppose that

$$\mathbb{E}(y_{it}|x_i) = \varphi(x_{it}\theta, \eta_i)$$

with φ strictly increasing in its first argument.

Then

$$\text{sign}\{\mathbb{E}(\Delta y_i|x_i)\} = \text{sign}\{\Delta x_i\theta\}$$

and the same estimator can be used.

Point identification requires a large-support condition in the regressors.

For any other value $\theta' \neq \theta$ we need that there exist (small set of) values x_i such that

$$\text{sign}(\Delta x_i\theta) \neq \text{sign}(\Delta x_i\theta')$$

happens with positive probability. Asking that Δx_i varies over the whole real line does this.

Bounds

Set identification

Consider a static problem.

The likelihood contribution of stratum i is

$$p(y_i|x_i, \theta, \eta_i) = \prod_{t=1}^{T} f(y_{it}|x_{it}, \theta, \eta_i)$$

and this is known up to finite-dimensional parameters θ, η_i.

The probabilities

$$p(y_i|x_i) = \int p(y_i|x_i, \theta, \eta_i)\, dG(\eta_i|x_i)$$

are nonparametrically identified from the data.

The identified set are all $\{\theta, G(\cdot|\cdot)\}$ for which the above equality holds.

This is not a singleton, in general.

Initial-condition problem

In a dynamic problem the distribution of the initial condition provides another parameter.

Take a first-order Markov process. Then

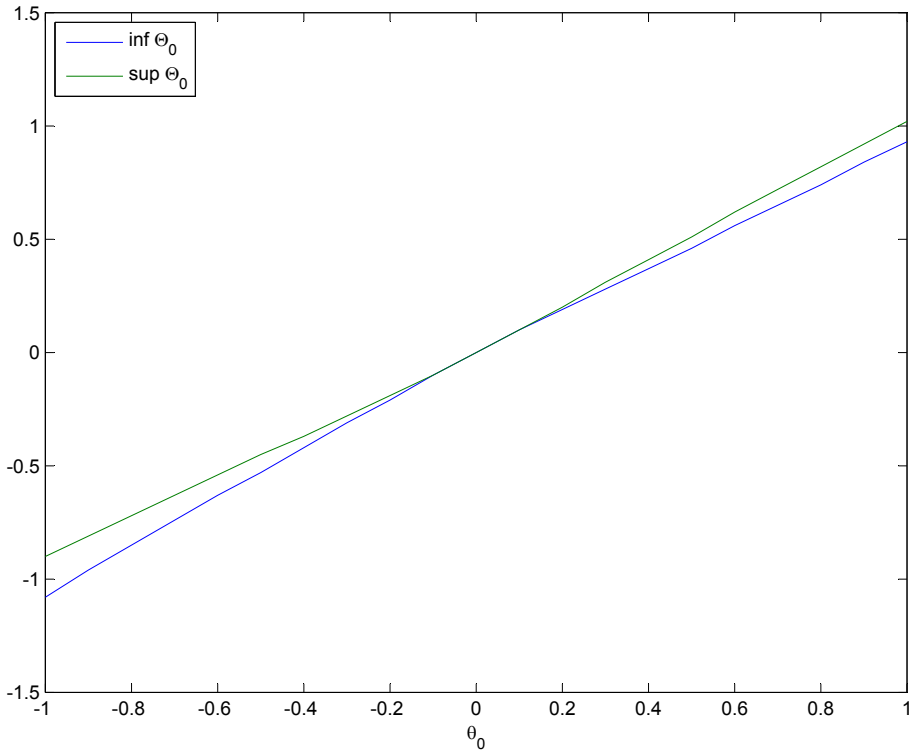
$$p(y_i|x_i, \theta, \eta_i, \pi(\cdot|\cdot)) = \prod_{t=1}^{T} f(y_{it}|, y_{it-1}, x_{it}, \theta, \eta_i)\, \pi(y_{i0}|x_i, \eta_i)$$

which features the additional unknown $\pi(\cdot|\cdot)$.

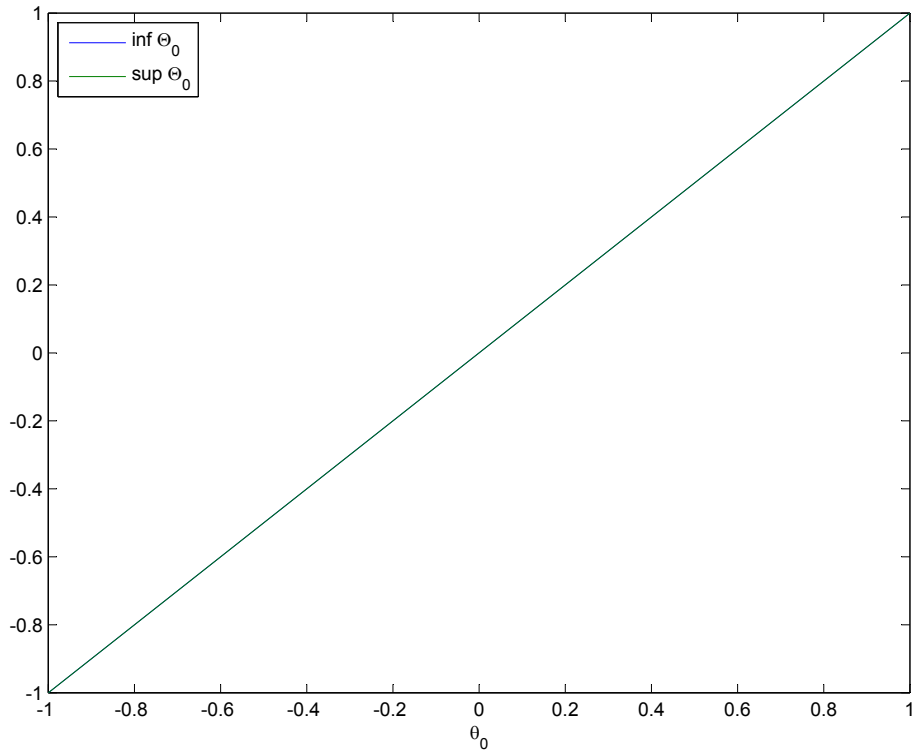
Examples

Autoregressive logit model with two transitions, with $P(y_{i0} = 1|\eta_i) = 1/2$ and η_i approx. normal (discretized).

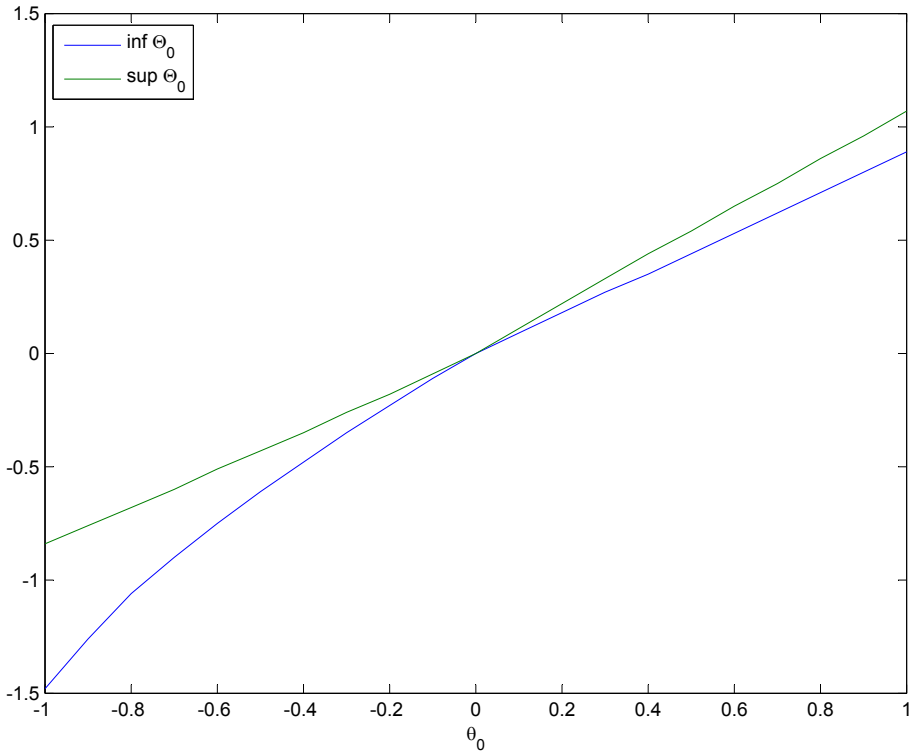
Identified set for θ:



Now adding a transition, the identified set for θ is a singleton:



Not so in the probit case:



Outline

1 Short panels
- Examples
- Recentering the score equation
- Conditional likelihood
- Ad hoc moment restrictions
- Bounds

2 Rectangular-array asymptotics
- Bias correction
- Parametric bootstrap
- Computation
- Average effects

Bias correction

Rectangular-array asymptotics

Potential for identification in short panels is limited.

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

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

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

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

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

An example: Linear model

Again,

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

Remember that, here,

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

has bias and variance

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

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

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

is not centered at zero.

The estimator is consistent but asymptotically biased.

Bias-correcting the estimator

In regular cases

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

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 a 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 $\hat{\theta}$.

Bias-correcting the profile likelihood

The profile log-likelihood is

$$\sum_{i=1}^{N} \ell_i(\theta, \hat{\eta}_i(\theta))$$

where

$$\hat{\eta}_i(\theta) = \arg\max_{\eta} \sum_{t=1}^{T} \log f(z_{it}|\theta, \eta).$$

It is a plug-in version of the infeasible target log-likelihood

$$\sum_{i=1}^{N} \ell_i(\theta, \eta_i(\theta))$$

where

$$\eta_i(\theta) = \arg\max_{\eta} \mathbb{E}(\log f(z_{it}|\theta, \eta)).$$

An expansion of

$$\ell_i(\theta, \hat{\eta}_i(\theta)) - \ell_i(\theta, \eta_i(\theta))$$

yields

$$\frac{\partial \ell_i(\theta, \eta_i(\theta))}{\partial \eta_i'} \left(\hat{\eta}_i(\theta) - \eta_i(\theta)\right) + \frac{1}{2}(\hat{\eta}_i(\theta) - \eta_i(\theta))' \frac{\partial^2 \ell_i(\theta, \eta_i(\theta))}{\partial \eta_i \partial \eta_i'} (\hat{\eta}_i(\theta) - \eta_i(\theta))$$

up to $o_p(T^{-3/2})$

We usually have

$$\hat{\eta}_i(\theta) - \eta_i(\theta) = \left(-\mathbb{E}\left(\frac{\partial^2 \ell_i(\theta, \eta_i(\theta))}{\partial \eta_i \partial \eta_i'}\right)\right)^{-1} \frac{\partial \ell_i(\theta, \eta_i(\theta))}{\partial \eta_i} + o_p(T^{-1/2}).$$

Pluging this in and taking expectations yields that the leading term of

$$\mathbb{E}(\ell_i(\theta, \hat{\eta}_i(\theta)) - \ell_i(\theta, \eta_i(\theta)))$$

equals

$$b_i(\theta) = \frac{1}{2}\text{trace}\left\{\mathbb{E}\left(-\frac{\partial^2 \ell_i(\theta, \eta_i(\theta))}{\partial \eta_i \partial \eta_i'}\right)^{-1} \mathbb{E}\left(\frac{\partial \ell_i(\theta, \eta_i(\theta))}{\partial \eta_i} \frac{\partial \ell_i(\theta, \eta_i(\theta))}{\partial \eta_i'}\right)\right\}.$$

The leading bias can be estimated by a sample version and subtracted from the profile log-likelihood:

$$\sum_{i=1}^{N} \hat{\ell}_i(\theta) - \hat{b}_i(\theta).$$

Under regularity conditions its maximizer has bias $o(T^{-1})$.

The first-order condition for this problem is

$$\sum_{i=1}^{N} \frac{\partial \hat{\ell}_i(\theta)}{\partial \theta} - \frac{\partial \hat{b}_i(\theta)}{\partial \theta} = 0.$$

The left-hand side here is a bias-corrected profile score.

Contrary to before, however, in the general case, it does not fully recenter the estimating equation at zero.

Other ways of approximately recentering the profile score are also possible.

Recall that the bias of the profile score

$$\sum_{i=1}^{N} \mathbb{E}_{\theta,\eta_i} \left(\frac{\partial \hat{\ell}_i(\theta)}{\partial \theta} \right) = \sum_{i=1}^{N} \mathbb{E}_{\theta,\eta_i} \left(\frac{\partial \ell_i(\theta, \hat{\eta}_i(\theta))}{\partial \theta} \right)$$

can always be computed for given θ and $\eta_1, \ldots, \eta_N$.

But we do not know which values for $\eta_1, \ldots, \eta_N$ to use.

For given θ, we could use $\hat{\eta}_i(\theta)$ and compute

$$\sum_{i=1}^{N} \mathbb{E}_{\theta,\hat{\eta}_i(\theta)} \left(\frac{\partial \hat{\ell}_i(\theta)}{\partial \theta} \right)$$

(by simulation, in general).

The adjusted estimating equation

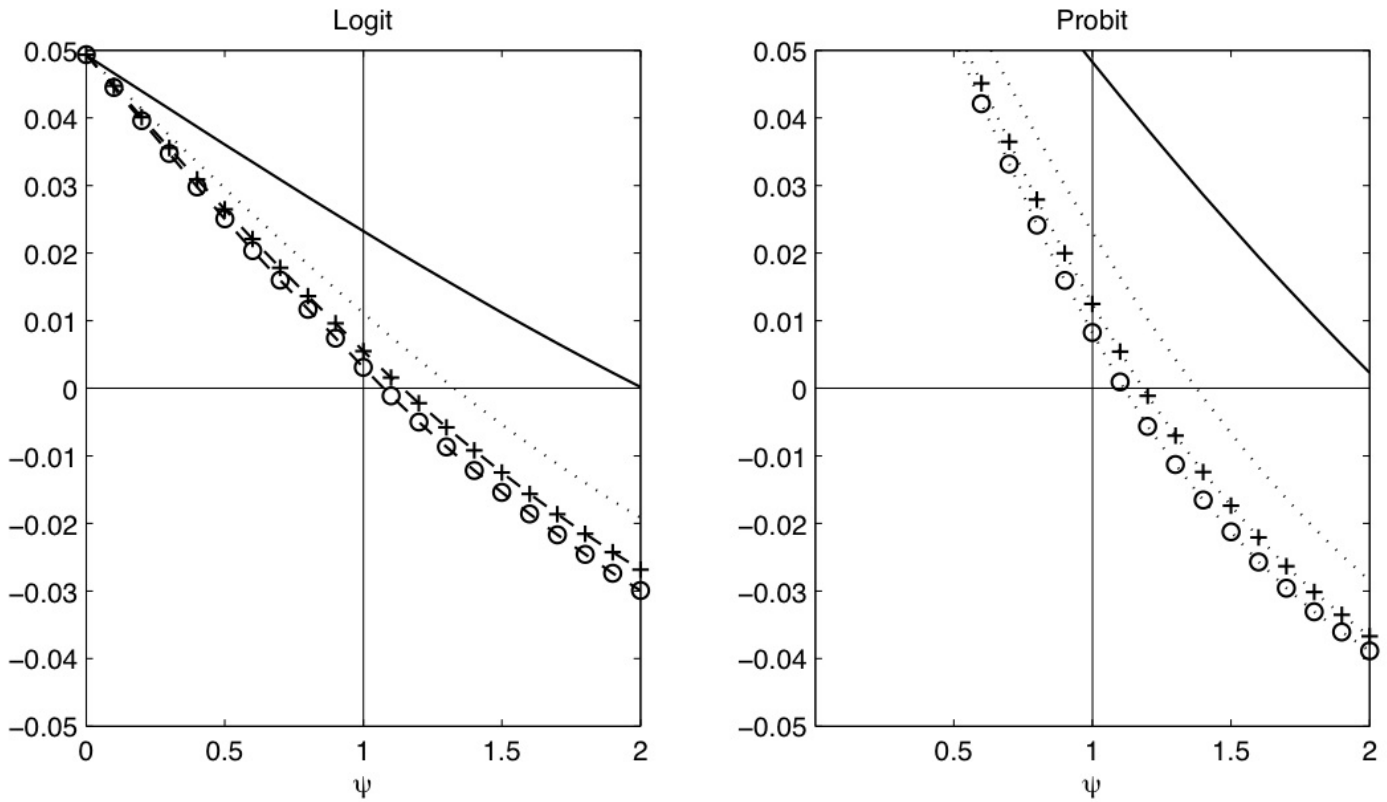
$$\sum_{i=1}^{N} \frac{\partial \hat{\ell}_i(\theta)}{\partial \theta} - \mathbb{E}_{\theta,\hat{\eta}_i(\theta)} \left(\frac{\partial \hat{\ell}_i(\theta)}{\partial \theta} \right) = 0$$

has bias $o(T^{-1})$.

This type of correction can be iterated to higher order.

An example: Logistic regression

Profile score and adjusted profile scores up to 3th order.



Simulations with standard-normal fixed effects, standard-normal regressor; $N = 250$.

	mean bias				median bias			
T	$\hat{\theta}$	$\dot{\theta}$	$\ddot{\theta}$	$\overline{\theta}$	$\widehat{\theta}$	$\dot{\theta}$	$\ddot{\theta}$	$\overline{\theta}$
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			
T	$\hat{\theta}$	$\dot{\theta}$	$\ddot{\theta}$	$\overline{\theta}$	$\widehat{\theta}$	$\dot{\theta}$	$\ddot{\theta}$	$\overline{\theta}$
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

Higher-order corrections

Suppose that

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

Then a second-order correction removes both B_1/T and B_2/T^2.

This translates into a correctly centered limit distribution as long as

$$N/T^3 \to c \in [0, \infty)$$

with $N, T \to \infty$.

Note that sequential estimation of the bias terms does not work, because it generates plug-in bias which alternates the expressions that appear in the expansion.

Joint estimation can be achieved through a jackknife.

To illustrate suppose that T is divisible by both 2 and 3.

As before first partition the panel into two subpanels to construct $\bar{\theta}_{1/2}$.

Next again partition the panel now into three subpanels to construct $\bar{\theta}_{1/3}$.

For the linear combination

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

to have zero bias up to second order we need that

$$\frac{1 + a_{1/2} + a_{1/3}}{T} - \frac{2a_{1/2}}{T} - \frac{3a_{1/3}}{T} = 0$$
$$\frac{1 + a_{1/2} + a_{1/3}}{T^2} - \frac{4a_{1/2}}{T^2} - \frac{9a_{1/3}}{T^2} = 0$$

which is a system of two equations in two unknowns, with solution $a_{1/2} = 3$, $a_{1/3} = -1$, leading to

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

having bias $o(T^{-2})$.

The iteration is now clear.

If

$$\text{plim}_{N\to\infty}\hat{\theta} = \theta + \frac{B_1}{T} + \frac{B_2}{T^2} + \cdots + \frac{B_k}{T^k} + o(T^{-k})$$

we can construct estimators whose bias is $o(T^{-k})$.

Iterating this to the limit will not typically yield consistency for fixed T, because of remainder terms.

Under rectangular-array asymptotics, all this is higher order.

One should also take into account variance considerations.

Bootstrap

The bootstrap

Given a parametric model we can simulate data for given parameter values.

Original data have density $f(z_{it}|\theta, \eta_i)$.

Maximum-likelihood estimator is

$$\hat{\theta}, \hat{\eta}_1, \ldots, \hat{\eta}_N = \arg\max \sum_{i=1}^{N} \sum_{t=1}^{T} \log f(z_{it}|\theta, \eta_i).$$

Simulate data from $f(z^*_{it}|\hat{\theta}, \hat{\eta}_i)$.

Maximum-likelihood estimator is

$$\hat{\theta}^*, \hat{\eta}^*_1, \ldots, \hat{\eta}^*_N = \arg\max \sum_{i=1}^{N} \sum_{t=1}^{T} \log f(z^*_{it}|\theta, \eta_i).$$

Under rectangular-array asymptotics this parametric bootstrap mimics the limit distribution of the maximum-likelihood estimator.

Valid confidence intervals can be constructed using the usual percentile method.

For example, an (equal-tailed) $1-\tau$ confidence interval for (scalar) θ is the set

$$\{\theta : \hat{\theta} - z^*_{1-\tau/2} \leq \theta \leq \hat{\theta} - z^*_{\tau/2}\}$$

where

$$z^*_\tau = \inf\{z^* : \tau \leq \mathbb{P}^*(\hat{\theta}^* - \hat{\theta} \leq z^*)\}$$

Here, we typically use simulation to approximate the bootstrap distribution.

This only requires a routine to compute the maximum-likelihood estimator.

No adjustment for bias needs to be made.

An example: Linear model

With

$$y_{it} \sim N(\eta_i, \theta)$$

we estimate $\hat{\eta}_i = \bar{y}_i$ and

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

Then simulate the distribution of the estimator by drawing bootstrap data as

$$y^*_{it} \sim N(\bar{y}_i, \hat{\theta}).$$

In this case

$$\sqrt{NT}(\hat{\theta} - \theta) \sim \text{Gamma}\left(-\sqrt{NT}\theta, \frac{N(T-1)}{2}, \frac{2\theta}{\sqrt{NT}}\right),$$

and, conditional on the original data,

$$\sqrt{NT}(\hat{\theta}^* - \hat{\theta}) \sim \text{Gamma}\left(-\sqrt{NT}\hat{\theta}, \frac{N(T-1)}{2}, \frac{2\hat{\theta}}{\sqrt{NT}}\right).$$

Because

$$\sqrt{NT}(\hat{\theta} - \theta) = -\sqrt{N/T}\,\theta + \epsilon$$

for mean-zero random variable $\epsilon = O_p(1)$, this is

$$\text{Gamma}\left(-\left(\sqrt{NT}\theta - \sqrt{\frac{N}{T}}\theta + \epsilon\right), \frac{N(T-1)}{2}, \frac{2\theta}{\sqrt{NT}}\left(1 - \frac{1}{T}\right) + \frac{2\epsilon}{NT}\right).$$

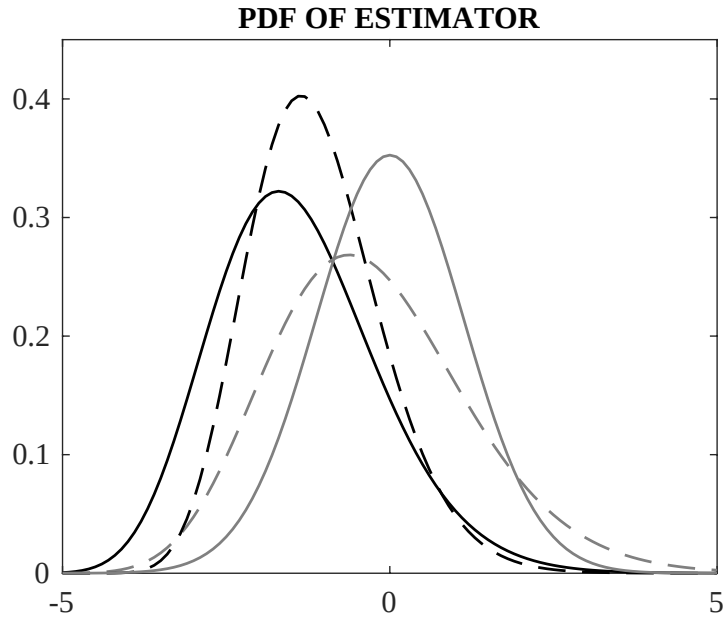
The mean and variance of the former distribution are

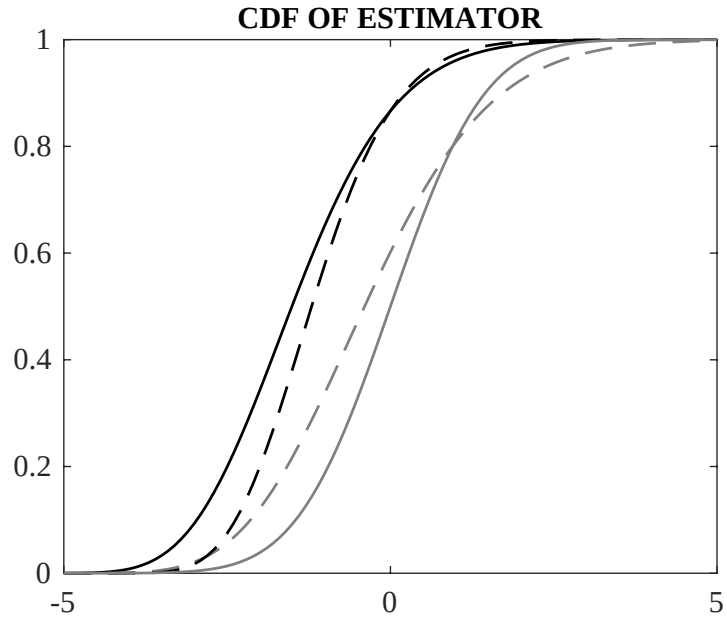
$$-\sqrt{\frac{N}{T}}\theta, \qquad 2\theta^2\left(1 - \frac{1}{T}\right).$$

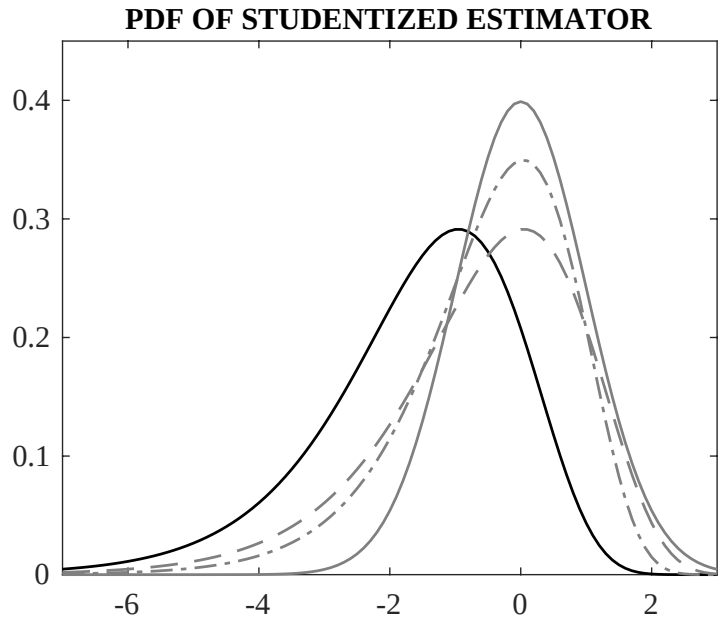
The mean and variance of the bootstrap distribution are

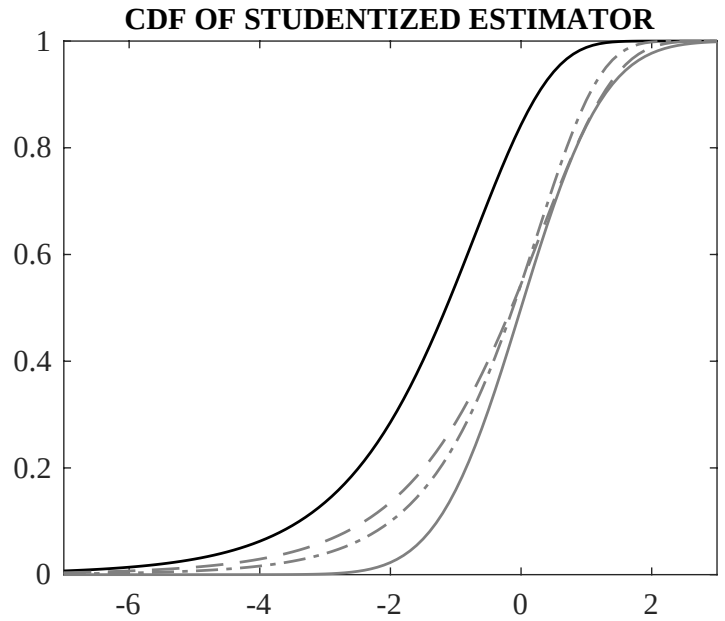
$$-\sqrt{\frac{N}{T}}\theta + \frac{1}{T}\left(\sqrt{\frac{N}{T}}\varphi_0 - \epsilon\right), \qquad 2\theta^2\left(1 - \frac{2}{T} + \frac{1}{T^2}\right) + O_p\left(\frac{1}{T}\right),$$

and these co-incide to first-order.


PDF OF ESTIMATOR
0.4
0.3
0.2
0.1
0
-5
0
5


CDF OF ESTIMATOR
1
0.8
0.6
0.4
0.2
0
-5
0
5


PDF OF STUDENTIZED ESTIMATOR
0.4
0.3
0.2
0.1
0
-6
-4
-2
0
2


CDF OF STUDENTIZED ESTIMATOR
1
0.8
0.6
0.4
0.2
0
-6
-4
-2
0
2

Computation

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-dimension matrices.

The score and Hessian are

$$\begin{pmatrix} \ell_\theta \\ \ell_{\eta_1} \\ \ell_{\eta_2} \\ \vdots \\ \ell_{\eta_N} \end{pmatrix}, \qquad \begin{pmatrix} \ell_{\theta\theta} & \ell_{\theta\eta_1} & \ell_{\theta\eta_2} & \cdots & \ell_{\theta\eta_N} \\ \ell_{\eta_1\theta} & \ell_{\eta_1\eta_1} & 0 & \cdots & 0 \\ \ell_{\eta_2\theta} & 0 & \ell_{\eta_2\eta_2} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \ell_{\eta_N\theta} & 0 & 0 & \cdots & \ell_{\eta_N\eta_N} \end{pmatrix},$$

where the individual components are

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

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

and

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

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

The inverse Hessian is

$$\ell^{-1} = \begin{pmatrix} (\ell^{-1})_{\theta\theta} & (\ell^{-1})_{\theta\eta_1} & (\ell^{-1})_{\theta\eta_2} & \cdots & (\ell^{-1})_{\theta\eta_N} \\ (\ell^{-1})_{\eta_1\theta} & (\ell^{-1})_{\eta_1\eta_1} & (\ell^{-1})_{\eta_1\eta_2} & \cdots & (\ell^{-1})_{\eta_1\eta_N} \\ (\ell^{-1})_{\eta_2\theta} & (\ell^{-1})_{\eta_2\eta_1} & (\ell^{-1})_{\eta_2\eta_2} & \ddots & (\ell^{-1})_{\eta_2\eta_N} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ (\ell^{-1})_{\eta_N\theta} & (\ell^{-1})_{\eta_N\eta_1} & (\ell^{-1})_{\eta_N\eta_2} & \cdots & (\ell^{-1})_{\eta_N\eta_N} \end{pmatrix},$$

with

$$\begin{aligned} (\ell^{-1})_{\theta\theta} &= \left(\ell_{\theta\theta} - \sum_{i=1}^{N} \ell_{\theta\eta_i}\, \ell^{-1}_{\eta_i\eta_i}\, \ell_{\eta_i\theta} \right)^{-1}, \\ (\ell^{-1})_{\theta\eta_i} &= -(\ell^{-1})_{\theta\theta} \ell_{\theta\eta_i} \ell^{-1}_{\eta_i\eta_i} \\ (\ell^{-1})_{\eta_i\eta_i} &= \ell^{-1}_{\eta_i\eta_i} + \ell^{-1}_{\eta_i\eta_i}\, \ell_{\eta_i\theta}\, (\ell^{-1})_{\theta\theta}\, \ell_{\theta\eta_i}\, \ell^{-1}_{\eta_i\eta_i} \\ (\ell^{-1})_{\eta_i\eta_j} &= \ell^{-1}_{\eta_i\eta_i}\, \ell_{\eta_i\theta}\, (\ell^{-1})_{\theta\theta}\, \ell_{\theta\eta_j}\, \ell^{-1}_{\eta_j\eta_j}. \end{aligned}$$

A Newton step for θ then is

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

A Newton step for η_i then is

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

Average effects

Parameters involving averages over fixed effects

In nonlinear models marginal effects are functions of the fixed effects.

In the binary choice model

$$\mathbb{P}(y_{it} = 1|x_i) = F(\eta_i + x_{it}\theta),$$

the effect of a marginal change in a continuous regressor is

$$\theta f(\eta_i + x_{it}\theta),$$

while the effect of a unit change in a discrete regressor x_{it} is

$$F(\eta_i + x_{it}\theta + \theta) - F(\eta_i + x_{it}\theta).$$

These objects vary with regressor values and individual effects.

Summary measures are averages over individual effects. Their (population) definition and estimation properties depend on how we treat the individual effects.

An example: Moments

No regressors.

Suppose we care about the average of $\varphi(\eta_i)$.

If we treat the η_i as parameters then the average of interest is

$$\frac{1}{N}\sum_{i=1}^{N}\varphi(\eta_i)$$

(Note how this changes with the sample size).

If we treat the η_i as independent draws from some distribution G then the average of interest is

$$\mathbb{E}(\varphi(\eta_i)) = \int \varphi(\eta)\, dG(\eta).$$

In both cases, we use

$$\frac{1}{N}\sum_{i=1}^{N}\varphi(\hat{\eta}_i)$$

as estimator.

Sampling behavior of both is different.

Consider

$$\hat{\eta}_i - \eta_i = \frac{\beta_i}{T} + \frac{1}{T}\sum_{t=1}^{T} \psi_{it} + o_p(T^{-1/2})$$

and suppose ψ_{it} is i.i.d.$(0, \sigma_i^2)$ over time for simplicity.

In the first sampling case all randomness comes from estimation of the effects. By a Taylor expansion

$$\frac{1}{N}\sum_{i=1}^{N} \varphi(\hat{\eta}_i) - \frac{1}{N}\sum_{i=1}^{N} \varphi(\eta_i)$$

equals

$$\frac{1}{N}\sum_{i=1}^{N} \left(\varphi'(\eta_i)\left(\hat{\eta}_i - \eta_i\right) + \frac{1}{2}\varphi''(\eta_i)\left(\hat{\eta}_i - \eta_i\right)^2 + o_p(T^{-1}) \right).$$

Then the bias of the plug-in estimator is

$$\frac{{}^1\!/_N \sum_{i=1}^N \varphi'(\eta_i)\beta_i}{T} + \frac{{}^1\!/_N \sum_{i=1}^N \varphi''(\eta_i)\sigma_i^2/T}{T},$$

which is of order T^{-1}.

Its first-order variance is equal to the variance of

$$\frac{1}{N}\sum_{i=1}^N \varphi'(\eta_i)\frac{1}{T}\sum_{t=1}^T \psi_{it}$$

and equals

$$\frac{1}{NT}\left({}^1\!/_N \sum_{i=1}^N \varphi'(\eta_i)^2\sigma_i^2\right)$$

so that the standard deviation is of order $(NT)^{-1/2}$.

Consequently, here, we have the same type of asymptotic bias problem as for common parameters.

In the second sampling case the bias is again of order T^{-1} (in the above expression it suffices to replace sample averages by expectations).

The variance, however, is of order N^{-1}. To see this consider the infeasible estimator

$$\frac{1}{N}\sum_{i=1}^{N}\varphi(\eta_i)$$

which is unbiased and $N^{-\frac{1}{2}}$ consistent for $\mathbb{E}(\varphi(\eta_i))$. It follows that

$$\begin{aligned}
{}^{1}/_{N}\sum_{i=1}^{N}\varphi(\hat{\eta}_i)-\mathbb{E}(\varphi(\eta_i)) &= \left({}^{1}/_{N}\sum_{i=1}^{N}\varphi(\hat{\eta}_i)-\varphi(\eta_i)\right)\\
&\quad+\left({}^{1}/_{N}\sum_{i=1}^{N}\varphi(\eta_i)-\mathbb{E}(\varphi(\eta_i))\right)\\
&= O\left(\frac{1}{T}\right)+O_p\left(\frac{1}{\sqrt{NT}}\right)+O_p\left(\frac{1}{\sqrt{N}}\right)
\end{aligned}$$

and so

$$\sqrt{N}\left({}^{1}/_{N}\sum_{i=1}^{N}\varphi(\hat{\eta}_i)-\mathbb{E}(\varphi(\eta_i))\right)\xrightarrow{d}N(0,\operatorname{var}(\varphi(\eta_i)))$$

as long as ${}^{N}/_{T^2}\to 0$.

Estimating the effects is irrelevant under rectangular-array asymptotics.

Estimation and inference

Note, though, that, in typical applications:

- bias correction is nevertheless useful, and
- the large-sample variance can yield a poor approximation.

A cross-sectional nonparametric bootstrap can be useful for inference here.

Analytical or jackknife bias correction is again possible.

In the first fixed-effect sampling design, resampling based on the parametric (recursive) bootstrap will also again be effective in mimicing the behavior of the estimator under rectangular-array asymptotics.