Second-order corrected likelihood for nonlinear fixed-effect models

Yutao Sun
EUR and NENU

Geert Dhaene
KU Leuven

June 2018

Introduction

Incidental parameter problem (Neyman & Scott 1948)

- MLE in short panels with fixed effects is often inconsistent
- Nickell (1981) bias is an example
- bias/inconsistency is generally $O(T^{-1})$

Exact vs. approximate solutions

- exact: works for specific models (GMM, conditional likelihood, marginal likelihood, invariant likelihood, functional differencing)
- approximate: works more generally (analytical, jackknife)

Approximate analytical solutions

- based on approximate bias formula for MLE or likelihood (Hahn & Newey 2004; Hahn & Kuersteiner 2011; Arellano & Hahn 2016; large statistics literature)
- current methods: first order, with bias $O(T^{-2})$
- this paper: second order, with bias $O(T^{-3})$; likelihood correction

Framework: panel data with incidental parameters

Parametric pdf

$$f(y_{it}|x_{it};\theta,\alpha_i) \qquad i=1,...,N;\ t=1,...,T$$

- θ is the parameter of interest
- $\alpha_1,...,\alpha_N$ are "incidental" parameters, "fixed effects"
- independence across i and t (restrictive)
- micropanels: N large, T small
- general (but known) f

Framework: panel data with incidental parameters

Incidental parameter problem

- when T is fixed, MLE is often inconsistent

$$\text{plim}_{N\to\infty}\widehat{\theta} \neq \theta \qquad \text{(with exceptions)}$$

- as $N, T \to \infty$ with $N/T \to c$,

$$\sqrt{NT}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(\text{bias}, V), \qquad \text{bias} \propto \sqrt{c}$$

Classic example: many normal means

$$\begin{aligned}
y_{it} &\sim \mathcal{N}(\alpha_i, \theta) \\
\widehat{\theta} &= \arg\max_{\theta} \left(-\frac{1}{2}\log\theta - \frac{1}{NT}\sum_{i,t}\frac{(y_{it} - \overline{y}_{i\cdot})^2}{2\theta} \right) \\
&= \frac{1}{NT}\sum_{i,t}(y_{it} - \overline{y}_{i\cdot})^2 \\
\text{plim}_{N\to\infty}\widehat{\theta} &= \theta - \theta/T
\end{aligned}$$

Correction

Let $\mathbb{E}$ be the expectation over (y_{it}, x_{it}). Let

$$\begin{aligned} L(\theta) &= \frac{1}{NT}\sum_{i,t}\log f\left(y_{it}|x_{it};\theta,\alpha_i(\theta)\right) \\ \alpha_i(\theta) &= \arg\max_{\alpha_i}\frac{1}{T}\sum_t \mathbb{E}\log f\left(y_{it}|x_{it};\theta,\alpha_i\right) \end{aligned}$$

Then for fixed T

$$\theta = \operatorname*{plim}_{N\to\infty}\arg\max_{\theta} L(\theta)$$

Think of $L(\theta)$ as an (infeasible) target log-likelihood to which an approximation may be constructed.

Correction

Let $\widehat{L}(\theta)$ be the concentrated (and normalized) loglikelihood

$$\begin{aligned}\widehat{L}(\theta) &= \frac{1}{NT}\sum_{i,t}\log f\left(y_{it}|x_{it};\theta,\widehat{\alpha}_i(\theta)\right) \\ \widehat{\alpha}_i(\theta) &= \arg\max_{\alpha_i}\frac{1}{T}\sum_{t}\log f\left(y_{it}|x_{it};\theta,\alpha_i\right)\end{aligned}$$

The difference $\widehat{L}(\theta) - L(\theta)$ can be expanded and approximated.

The kth-order approximation to $L(\theta)$ takes the form

$$\widehat{L}^{(k)}(\theta) = \widehat{L}(\theta) + \frac{\widehat{B}_1(\theta)}{T} + \frac{\widehat{B}_2(\theta)}{T^2} + \cdots + \frac{\widehat{B}_k(\theta)}{T^k}$$

where $\widehat{B}_j(\theta)$ depends on the $\widehat{\alpha}_i(\theta)$ (not the $\alpha_i(\theta)$) and satisfies

$$\mathbb{E}\widehat{L}^{(k)}(\theta) - \mathbb{E}L(\theta) = O(T^{-k-1})$$

(or with $\operatorname{plim}_{N\to\infty}$ instead of $\mathbb{E}$). Then we can take

$$\widehat{\theta}^{(k)} = \arg\max_{\theta}\widehat{L}^{(k)}(\theta)$$

Correction

For the uncorrected likelihood

$$\mathbb{E}\widehat{L}\left(\theta\right) - \mathbb{E}L\left(\theta\right) = O(T^{-1})$$

Arellano & Hahn 2016 derive $\widehat{B}_1\left(\theta\right)$, giving $\widehat{L}^{(1)}\left(\theta\right)$ satisfying

$$\mathbb{E}\widehat{L}^{(1)}\left(\theta\right) - \mathbb{E}L(\theta) = O(T^{-2})$$

We derive $\widehat{B}_2\left(\theta\right)$, giving $\widehat{L}^{(2)}\left(\theta\right)$ satisfying

$$\mathbb{E}\widehat{L}^{(2)}\left(\theta\right) - \mathbb{E}L(\theta) = O(T^{-3})$$

Correction

The correction terms are

$$\widehat{B}_1(\theta) = \frac{1}{N}\sum_i \widehat{b}_{1i}(\theta), \qquad \widehat{B}_2(\theta) = \frac{1}{N}\sum_i \widehat{b}_{2i}(\theta),$$

where, omitting i,

$$\widehat{b}_1(\theta) = \frac{\widehat{\mathcal{L}}_1\binom{2}{1}}{2\widehat{l}_2} \qquad \text{(as in Arellano \& Hahn 2016)}$$

$$\begin{aligned}\widehat{b}_2(\theta) = & -\frac{\widehat{\mathcal{L}}_2\binom{1,1;1,1}{1,2;1,2}}{\widehat{l}_2^3} - \frac{\widehat{l}_3\widehat{\mathcal{L}}_1\binom{3}{1}}{3\widehat{l}_2^3} - \frac{\widehat{l}_4\widehat{\mathcal{L}}_2\binom{2;2}{1;1}}{12\widehat{l}_2^4} + \frac{5\widehat{l}_4\widehat{\mathcal{L}}_1\binom{2}{1}^2}{24\widehat{l}_2^4} \\ & -\frac{5\widehat{l}_3^2\widehat{\mathcal{L}}_1\binom{2}{1}^2}{8\widehat{l}_2^5} + \frac{\widehat{l}_3^2\widehat{\mathcal{L}}_2\binom{2;2}{1;1}}{4\widehat{l}_2^5} + \frac{\widehat{\mathcal{L}}_1\binom{2,1}{1,2}}{\widehat{l}_2^2} - \frac{\widehat{\mathcal{L}}_1\binom{2}{1}\widehat{\mathcal{L}}_1\binom{2}{2}}{2\widehat{l}_2^3} \\ & -\frac{\widehat{\mathcal{L}}_1\binom{2}{1}\widehat{\mathcal{L}}_1\binom{1,1}{1,3}}{2\widehat{l}_2^3} + \frac{3\widehat{l}_3\widehat{\mathcal{L}}_1\binom{2}{1}\widehat{\mathcal{L}}_1\binom{1,1}{1,2}}{2\widehat{l}_2^4}\end{aligned}$$

$\widehat{l}_r$ and $\widehat{\mathcal{L}}_{\mathcal{P}}$ are sums of products of loglikelihood derivatives wrt α_i, to be evaluated at $\alpha_i = \widehat{\alpha}_i(\theta)$.

Correction

l_r and $\mathcal{L}_{\mathcal{P}}$

$$l_r = \frac{1}{T}\sum\nolimits_t \left\{\nabla^r_{\alpha_i} \log f\left(y_{it}|x_{it};\theta,\alpha_i\right)\right\}_{\alpha_i=\alpha_i(\theta)}$$

$$\begin{aligned}&\mathcal{L}_{\mathcal{P}}\binom{p_{11},\cdots,p_{1M};\cdots;p_{J1},\cdots p_{JM}}{r_{11},\cdots,r_{1M};\cdots;r_{J1},\cdots r_{JM}}\\ &\qquad = \frac{1}{T^{\mathcal{P}}}\sum_{t_1,\cdots,t_J}\prod_{j=1}^{J}\prod_{m=1}^{M}\left\{\nabla^{r_{jm}}_{\alpha_i}\log f\left(y_{it_j}|x_{it_j};\theta,\alpha_i\right)\right\}^{p_{jm}}_{\alpha_i=\alpha_i(\theta)}\end{aligned}$$

with all $t_1,\cdots,t_J$ distinct and $\mathcal{P}$ chosen such that

$$\mathcal{L}_{\mathcal{P}}\binom{p_{11},\cdots,p_{1M};\cdots;p_{J1},\cdots p_{JM}}{r_{11},\cdots,r_{1M};\cdots;r_{J1},\cdots r_{JM}} = O_p(1)$$

Example

$$\mathcal{L}_1\binom{2,1}{1,2} = \frac{1}{T}\sum\nolimits_t \left\{\left(\nabla_{\alpha_i}\log f\left(y_{it}|x_{it};\theta,\alpha_i\right)\right)^2 \nabla^2_{\alpha_i}\log f\left(y_{it}|x_{it};\theta,\alpha_i\right)\right\}_{\alpha_i=\alpha_i(\theta)}$$

Example: many normal means

Second-order correction

$$\widehat{L}^{(2)}(\theta) = -\frac{1}{2}\log\theta - \frac{1}{N}\left(\frac{1}{T}+\frac{1}{T^2}+\frac{1}{T^3}\right)\sum_{i,t}\frac{(y_{it}-\overline{y}_{i\cdot})^2}{2\theta}$$

$$\widehat{\theta}^{(2)} = \frac{1}{N}\left(\frac{1}{T}+\frac{1}{T^2}+\frac{1}{T^3}\right)\sum_{i,t}(y_{it}-\overline{y}_{i\cdot})^2$$

$$\text{plim}_{N\to\infty}\widehat{\theta}^{(2)} = \theta - \theta/T^3$$

Compare with first-order correction (Arellano & Hahn 2016)

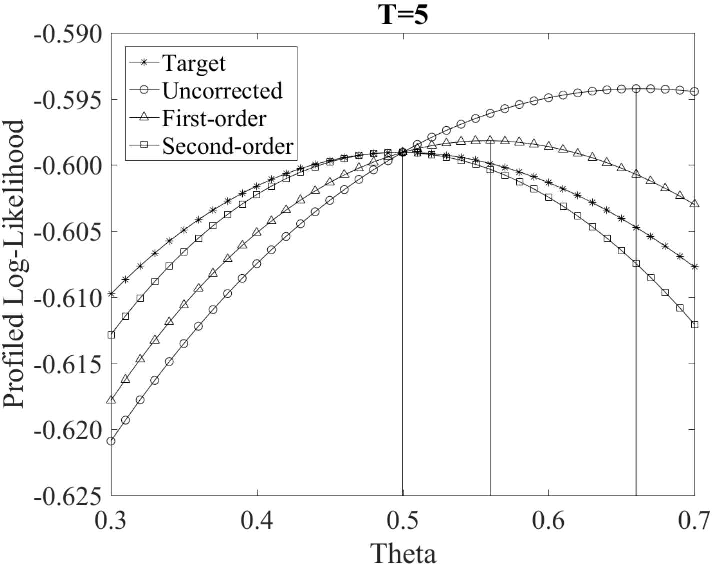
$$\widehat{L}^{(1)}(\theta) = -\frac{1}{2}\log\theta - \frac{1}{N}\left(\frac{1}{T}+\frac{1}{T^2}\right)\sum_{i,t}\frac{(y_{it}-\overline{y}_{i\cdot})^2}{2\theta}$$

$$\widehat{\theta}^{(1)} = \frac{1}{N}\left(\frac{1}{T}+\frac{1}{T^2}\right)\sum_{i,t}(y_{it}-\overline{y}_{i\cdot})^2$$

$$\text{plim}_{N\to\infty}\widehat{\theta}^{(1)} = \theta - \theta/T^2$$

Example: logit

Single data set with $T = 5, N = 10,000, x_{it} \sim \mathcal{N}(0, 4)$, $\alpha_i = 0$, $\theta = 0.5$. Curves are vertically shifted to coincide at θ.



Example: logit

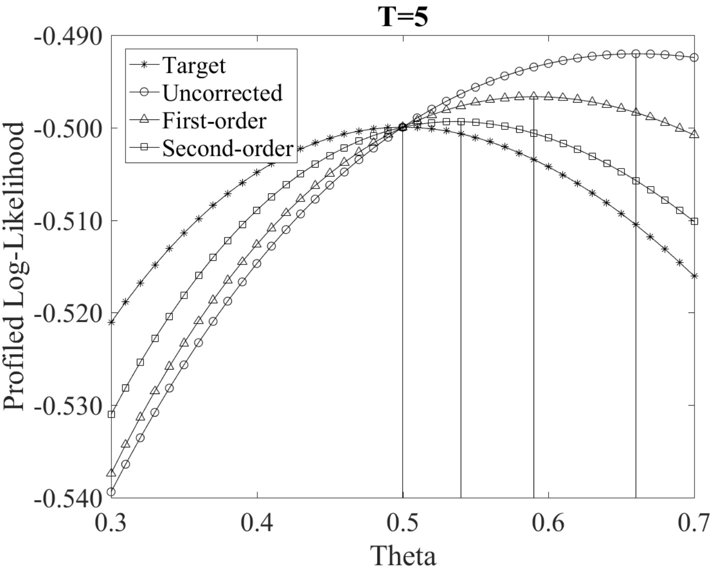
1000 simulations with $N = 10,000, x_{it} \sim \mathcal{N}(0,1)$, $\alpha_i = 0$

	Mean	Bias	RMSE
$T = 5$		$\theta = 0.5$	
$\widehat{\theta}$	0.6310	0.1310	0.1318
$\widehat{\theta}^{(1)}$	0.5379	0.0379	0.0396
$\widehat{\theta}^{(2)}$	0.4910	$-\mathbf{0.0090}$	0.0138
$T = 10$		$\theta = 0.5$	
$\widehat{\theta}$	0.5570	0.0570	0.0576
$\widehat{\theta}^{(1)}$	0.5092	0.0092	0.0120
$\widehat{\theta}^{(2)}$	0.4990	$-\mathbf{0.0010}$	0.0076

order of bias: $\mathbf{0.0010}/\mathbf{0.0090} \approx 1/8$, as predicted

Example: probit

Single data set with $T = 5, N = 10,000, x_{it} \sim \mathcal{N}(0, 4)$, $\alpha_i = 0$, $\theta = 0.5$. Curves are vertically shifted to coincide at θ.



Example: probit

1000 simulations with $N = 10,000, x_{it} \sim \mathcal{N}(0,1)$, $\alpha_i = 0$

	Mean	Bias	RMSE
$T = 5$		$\theta_0 = 0.5$	
$\widehat{\theta}$	0.6239	0.1239	0.1242
$\widehat{\theta}^{(1)}$	0.5581	0.0581	0.0587
$\widehat{\theta}^{(2)}$	0.5180	0.0180	0.0195
$T = 10$		$\theta_0 = 0.5$	
$\widehat{\theta}$	0.5504	0.0504	0.0507
$\widehat{\theta}^{(1)}$	0.5121	0.0121	0.0130
$\widehat{\theta}^{(2)}$	0.5005	0.0005	0.0047

Likelihood ratio test: probit model

10,000 simulations with $N = 100, x_{it} \sim \mathcal{N}(0,1)$, $\alpha_i \sim \mathcal{N}(0, 1/16)$, $\theta = 0.5$

Table: Rejection rates of 5%-level LR test

H_0	0.48	0.49	0.5	0.51	0.52
			$T = 5$		
$\widehat{L}(\theta)$	1.00	1.00	1.00	1.00	0.99
$\widehat{L}^{(1)}(\theta)$	0.96	0.90	0.81	0.68	0.53
$\widehat{L}^{(2)}(\theta)$	0.57	0.40	0.25	0.14	0.09
$L(\theta)$	0.27	0.12	0.05	0.05	0.12
			$T = 10$		
$\widehat{L}(\theta)$	0.99	0.97	0.90	0.76	0.54
$\widehat{L}^{(1)}(\theta)$	0.62	0.37	0.18	0.07	0.08
$\widehat{L}^{(2)}(\theta)$	0.33	0.15	0.06	0.10	0.25
$L(\theta)$	0.32	0.12	0.05	0.10	0.28

Outline of derivation

Let

$$\begin{aligned} \widehat{l} &= \widehat{l}\,(\theta) = \frac{1}{T}\sum_t \log f\left(y_{it}|x_{it};\theta,\widehat{\alpha}_i\left(\theta\right)\right) \\ l &= l\left(\theta\right) = \frac{1}{T}\sum_t \log f\left(y_{it}|x_{it};\theta,\alpha_i\left(\theta\right)\right) \end{aligned}$$

Expand $\widehat{l}_1 = 0$ around $\alpha_i\left(\theta\right)$ and rearrange to obtain

$$\widehat{\alpha}_i\left(\theta\right) - \alpha_i\left(\theta\right) = -\frac{l_1}{l_2} - \frac{l_1^2 l_3}{2l_2^3} + \frac{l_1^3 l_4}{6l_2^4} - \frac{l_3^2 l_1^3}{2l_2^5} + O_p(T^{-2})$$

(similar to Rilstone, Srivastava & Ullah 1996)

Outline of derivation

Expand $\widehat{l}$ around $\alpha_i(\theta)$ to obtain

$$\begin{aligned} l &= \widehat{l} - l_1\left(\widehat{\alpha}_i(\theta) - \alpha_i(\theta)\right) - \frac{1}{2} l_2 \left(\widehat{\alpha}_i(\theta) - \alpha_i(\theta)\right)^2 \\ &\quad - \frac{1}{6} l_3 \left(\widehat{\alpha}_i(\theta) - \alpha_i(\theta)\right)^3 - \frac{1}{24} l_4 \left(\widehat{\alpha}_i(\theta) - \alpha_i(\theta)\right)^4 + O_p(T^{-5/2}) \end{aligned}$$

Use representation of $\widehat{\alpha}_i(\theta) - \alpha_i(\theta)$ and group terms to give

$$\mathbb{E}l = \mathbb{E}\widehat{l} + \frac{\mathbb{E}b_1}{T} + \frac{\mathbb{E}b_2'}{T^2} + O(T^{-3})$$

where b_1 and b_2' are evaluated at $\alpha_i(\theta)$.

Outline of derivation

Replacing b_1 with $\widehat{b}_1$ (b_1 evaluated at $\widehat{\alpha}_i(\theta)$) introduces a bias

$$\mathbb{E}b_1 = \mathbb{E}\widehat{b}_1 + O\left(T^{-1}\right)$$

so

$$\mathbb{E}l = \mathbb{E}\widehat{l} + \frac{\mathbb{E}\widehat{b}_1}{T} + \frac{\mathbb{E}b_2'}{T^2} + O(T^{-2})$$

Replace $(l, \widehat{l})$ by $(b_1, \widehat{b}_1)$ in the earlier procedure to obtain

$$\frac{\mathbb{E}b_1}{T} = \frac{\mathbb{E}\widehat{b}_1}{T} + \frac{\mathbb{E}b_{1,1}}{T^2} + O(T^{-3})$$

Rearrange.

Concluding remarks

- Correcting the likelihood is easier than correcting the MLE
- Asymptotic refinement. But in general it does not lead to fixed-T consistency
- Comparison with jackknife corrections
  - effect on higher-order bias terms is different (and more easy to analyze for jackknife)
  - computational: "delete-one-delete-two" jackknife is more intensive; split-panel jackknife is less intensive
- Comparison with profile score bias corrections (McCullagh & Tibshirani 1990)
  - profile score bias calculation is a fixed-T calculation and the correction is exact whenever the bias is free of α_i
  - computational: profile score bias calculation may require numerical integration
- Extension to arbitrary order?
- Extension to dynamic models?