Missing Values in Linear Dynamic Panel Models

Dr. Marcel Preising
Federal Statistical Office of Germany

September 2020

- Linear dynamic panel regression models are commonly used tools in economics, social sciences and anthropology, e.g. for
  - explaining dynamics in economic growth,
  - ...

- Estimation routines so far require completely observed data sets.

$\Longrightarrow$ Common problem in empirical data: **Missing values!**

- Objectives:
  - To deal with missing values in the (lagged) dependent and explanatory variables.
  - (To enable non-nested model comparison.)

Lagged dependent variable model of order one,
with fixed effects (FE) or random effects (RE):

$$y_{i,1} = c_i + a_1 + X_{i,1}\beta + \varepsilon_{i,1}, \qquad i = 1, \ldots, N,$$

and

$$y_{i,t} = c_i + a_t + X_{i,t}\beta + \rho y_{i,t-1} + \varepsilon_{i,t}, \quad i = 1, \ldots, N, \quad t = 2, \ldots, T,$$

c_i:	individual-specific effect (Fixed or random effects),
a_t:	time-specific effect (Fixed or random effects),
$X_{i,t}$:	covariate information,
$y_{i,t-1}$:	lagged dependent variable,
$\varepsilon_{i,t} \sim N(0, \sigma_t^2)$:	period-specific error variance.

If random effects: $c_i \sim N(0, \sigma_c^2)$, $a_t \sim N(0, \sigma_a^2)$.

Bayesian estimation algorithm:

$y^{(m)} = [y^{\text{obs}}, y^{\text{mis},(m)}]$ and $X^{(m)} = [X^{\text{obs}}, X^{\text{mis},(m)}]$, for $m = 1, \ldots, M$.
With initialization $\theta^{(0)}, y^{(0)}$, and $X^{(0)}$,

Step *I* Sample $\theta^{(m)}$ from $f(\theta | y^{(m-1)}, X^{(m-1)})$.

Step *II* Sample $y^{\text{mis},(m)}$ from $f(y^{\text{mis}} | y^{\text{obs}}, \theta^{(m)}, X^{(m-1)})$.

Step *III* Sample $X^{\text{mis},(m)}$ from $f(X^{\text{mis}} | y^{(m)}, \theta^{(m)}, X^{\text{obs}})$.

For Step I, formulate prior and posterior distributions:

- Conjugate prior distributions:
  FE:

$$\begin{aligned} \beta, \rho, \{c_i\}_{i=1}^{N-1}, \{a_t\}_{t=1}^{T-1} : &\quad \text{(Multivariate) Normal} \\ \{\sigma_t^2\}_{t=1}^{T} : &\quad \text{Inverse Gamma} \end{aligned}$$

  RE:

$$\begin{aligned} \beta, \rho : &\quad \text{(Multivariate) Normal} \\ \{\sigma_t^2\}_{t=1}^{T}, \sigma_c^2, \sigma_a^2 : &\quad \text{Inverse Gamma} \end{aligned}$$

- Full conditional posterior distributions allow for Gibbs-Sampling (e.g. Geman and Geman, 1984).

For Step *II*, reformulate

$$f(y|\theta, X) = \prod_{i=1}^{N} f(y_{i,2:T}|\theta, X_{i,2:T}, y_{i,1}) f(y_{i,1}|\theta, X_{i,1})$$

$\Longrightarrow$ Use properties of multivariate normal distribution to draw missing values. $\rightarrow$ Data Augmentation.

For Step *III*:

- Sequential Classification and Regression Trees (CART)-Algorithms with Bayesian Bootstrap. (Code from R-package *multiple imputation by chained equations*, van Buuren et al., 2020)
  - → Has shown to be highly capable of dealing with discrete and possible nonlinear relationships among the variables (Doove et al., 2014).

Approximated Log-marginal likelihood (Chib, 1995):

$$\ln \widehat{f}(y|X) = \ln f(y|\tilde{\theta}, X) + \ln f(\tilde{\theta}) - \ln \widehat{f}(\tilde{\theta}|y, X),$$

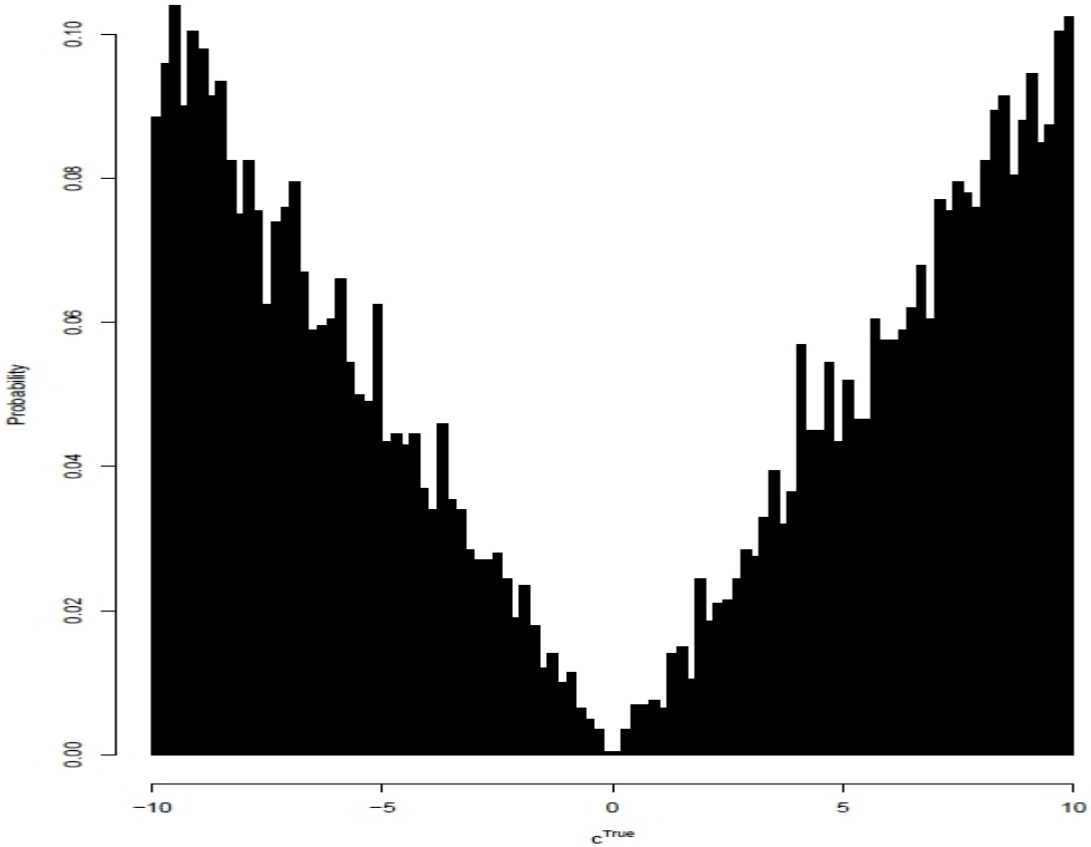
with $\tilde{\theta}$ denoting the posterior estimates resulting from the M Gibbs draws:

- Objectives:
  - Comparison of Data Augmentation and Complete Case.
  - (Reliable model comparison by marginal likelihoods.)

- 500 replications.
- Data-generating process:

  - N=50, T=5,
  - $X_{i,t} \sim \mathrm{N}(\mu_X = \mathbf{0}, \mathrm{corr}_X = \begin{pmatrix} 1 & 0.35 & 0.5 \\ 0.35 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix})$,
    and $\mathcal{I}(X_{3,i,t} < -1) + 2\mathcal{I}(-1 \leq X_{3,i,t} \leq 1) + 3\mathcal{I}(X_{3,i,t} > 1)$.

- $\beta = (-1.630, 4.900, 4.528, 1.437);\ \rho = 0.8.$
- RE: $\sigma_c^2 = \sigma_a^2 = 0.75.$
- FE: $f(c,a) = \prod_{i=1}^{N-1} \frac{|c_i|}{100} \mathcal{I}_{[-10,10]}(c_i) \prod_{t=1}^{T-1} \frac{|a_t|}{100} \mathcal{I}_{[-10,10]}(a_t)$:



- Three missing scenarios:

  $\Longrightarrow$ **Scenario** I**:**
  Before Deletion.

  $\Longrightarrow$ **Scenario** II**:**
  $\Pr(y_{it} \text{ is missing}) = \frac{1}{1+\exp\{2-2.5X_{2,i,t}\}} \approx$ **25**% Missing.

  $\Longrightarrow$ **Scenario** III**:**
  $X_{3,i,t}$ is missing if $F_U(U_{i,t}) > 0.9$, with $F_U(U_{i,t})$ as the empirical distribution function of the random variable $U_{i,t}$, and

  $$U_{i,t} = \frac{1}{1+\exp\{0.2\phi_{i,t}X_{1,i,t}\}+\tau_{i,t}}, \qquad i = 1,\ldots,N, \quad t = 1,\ldots,T,$$

  where both $\phi_{i,t}$ and $\tau_{i,t}$ are standard normally distributed.
  Thus, exactly **10**% set to be missing.

- Prior specifications:
  - Multivariate Normal: Expectation 0 / Variance 50.
  - Inverse Gamma: Shape 1 / Scale 1.

Table: Simulation Study – Results I:
Root mean squared errors and log-marginal likelihoods. DGP: Fixed effects model.

		DGP: FE					
		FE Analysis			**RE Analysis**		
		I	*II*	*III*	*I*	*II*	*III*
Data Augm.	β_2	0.121	0.151	0.153	0.122	0.152	0.154
	β_3	0.123	0.168	0.170	0.123	0.167	0.170
	β_4	0.225	0.287	**0.301**	0.226	0.287	0.303
	ρ	0.010	0.012	0.012	0.011	0.012	0.012
Compl.Case	β_2	–	0.188	0.210	–	0.191	0.212
	β_3	–	0.211	0.229	–	0.209	0.228
	β_4	–	0.343	**0.375**	–	0.345	0.376
	ρ	–	0.018	0.021	–	0.020	0.023
Preference rate (%):		**99.8**	**99.6**	**99.6**	0.2	0.4	0.4

Root mean squared errors. Preference rates indicate how often the specifications are chosen as the favorable analysis model in terms of the log-marginal likelihood.

Table: Simulation Study – Results II:
Root mean squared errors and log-marginal likelihoods. DGP: Random effects model.

		DGP: RE					
		FE Analysis			**RE Analysis**		
		I	*II*	*III*	*I*	*II*	*III*
Data Augm.	β_2	0.137	0.160	0.160	0.123	0.142	0.147
	β_3	0.136	0.177	0.177	0.121	0.158	0.161
	β_4	0.241	0.284	0.303	0.227	0.265	0.278
	ρ	0.024	0.031	0.032	0.013	0.016	0.016
Compl.Case	β_2	–	0.177	0.188	–	0.155	0.163
	β_3	–	0.231	0.253	–	0.193	0.208
	β_4	–	0.340	0.362	–	0.297	0.305
	ρ	–	0.031	0.032	–	0.019	0.019
Preference rate (%):		0	0	0	**100**	**100**	**100**

Root mean squared errors. Preference rates indicate how often the specifications are chosen as the favorable analysis model in terms of the log-marginal likelihood.

- Concluding remarks:
  - Efficient estimation due to data augmentation approach.
  - Reliable model comparison (also for incomplete data).

- Future Work:
  - Extension for binary response variables (in progess).
  - Applications for official statistics:
    - → **Challenge**: Imputation methods satisfying constraints (edit restrictions and preservation of totals).

Thank you for your attention!