Efficient GMM estimation with a general missing data pattern

Chris Müris

Simon Fraser University

December 2012

Missing data in economics

40% of empirical studies work with missing data

70% of those use a *complete case* estimator

Source: Donald and Abrevaya (2010)

Incomplete observations

- an observation is incomplete not all variables are observed
- incomplete observations can be informative

Examples

- missing instruments
- unbalanced panel data
- no gain: OLS

Efficiency

- Contribution: Propose an estimator
  - for a general GMM setting
  - that handles any missing data pattern
  - that is efficient
- Efficient estimator
  - uses all informative observations
  - combines them in an optimal way
- Substantial efficiency gains over naive estimators

Proposed estimator

1. Create subsamples based on missing data pattern
2. Determine optimal estimator for each subsample
3. Find the optimal combination of subsample estimators

Why you should use it

Properties

- Asymptotically efficient
- Easy to implement and computationally cheap
- Extendable (IPW, CUE, GEL)

Flexibility

- Any standard GMM setting with a random sample
- General *pattern* of missing data

Overview

1	Example, literature	
2	Model	*no selection*
3	Efficient estimation	
4	Examples	
5	Inverse probability weighting	*selection on observables*

Multiple instruments

- Regression model with endogenous variables
- Two instruments are available
- For some observations, only one (or none) is observed
- Examples: many! Levitt (2002), Rodrik et al. (2004)

(Dynamic) panel data

- $N \to \infty$, fixed-T
- Unit of observation is the individual/firm/country
- Incomplete if not all time periods are available
- Examples: Arellano and Bond (1991), Blundell and Bond (1998), Meghir and Pistaferri (2004)

Instrument example: Institutions or geography?

- What determines economic development?
- Acemoglu, Johnson and Robinson (2001) look at former European colonies
- Model: income (INC) depends on geography (GEOG) and institutions (INST)

$$\mathrm{INC}_i = \beta_0 + \beta_1 \mathrm{GEOG}_i + \beta_2 \mathrm{INST}_i + \varepsilon_i$$

Endogeneity

- Institutions are endogenous
- Contribution AJR: settler mortality rate (MORT) instrument
- Settler mortality rate
  - affected early settlement . . .
  - which affected early institutions. . .
  - which affected current institutions. . .
  - which affect current output.

Trade

- Rodrik, Subramanian and Trebbi (2004) incorporate trade
- Trade (TRADE) is endogenous
- Instrument: constructed share trade (SHARE) from Frankel and Romer (1999)

“Institutions rule”

- Institutions: significant
- Geography: not significant

Missings instruments

For any observation, we have

- no instrument, or
- one instrument, or
- both instruments.

Common practice

- complete case
- complete instruments
- available case
- generate instruments
- imputation
- ...

Theoretical literature

1. parametric approach: Little and Rubin (2002)
2. imputation: Little and Rubin (2002); Wang, Linton and Härdle (2004)
3. complete or uninformative: "all-or-nothing"
   Robins, Rotnitzky and Zhao (1994); Wooldridge (2007); Chen, Hong and Tarozzi (2008); Graham (2010)
4. specific non-all-or-nothing settings:
   Robins, Rotnitzky and Zhao (1995); Abowd, Crépon and Kramarz (2007); Lynch and Wachter (2008); Abrevaya and Donald (2010)

Contribution: any pattern is handled; no imputation required; semiparametric

Model: Overview

- Assume a standard GMM model
- Assume that it holds regardless of what data is available

Generalized method of moments

- We are interested in a parameter $\theta_0 \in \mathbb{R}^p$
- We know that moment condition $\mathbb{E}(h(X, \theta_0)) = 0$ holds
  - $X \in \mathbb{R}^d$ is a random vector
  - $h : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^q$ is a moment function
- A random sample for X is available
- Consider the sample moment $\bar{h}_n(\theta) = \frac{1}{n} \sum_{i=1}^n h(X_i, \theta)$
- GMM estimator minimizes a quadratic form of it

$$\hat{\theta}_n = \text{argmin}_\theta \, \bar{h}_n(\theta)' W_n \bar{h}_n(\theta)$$

Optimal GMM

- As the number of observations n grows,
- And we choose the weight matrix W_n optimally,
- The limiting distribution of optimal estimator is

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to N\left(0, \left(D'\Omega^{-1}D\right)^{-1}\right),$$

  - $D = \mathbb{E}\left(\frac{\partial h(X,\theta_0)}{\partial \theta'}\right)$ is the derivative at the truth
  - $\Omega = \mathrm{var}\left(h(X,\theta_0)\right)$ is the variance of h at the truth

Missing data pattern

- Usually, missing data indicator is $\{0, 1\}$
- Our generalized missing data indicator R is a random matrix
- Indicates which components of h we observe

Rodrik example (1)

- INC is log PPP GDP per capita in 1995
- GEOG is distance of capital city to equator
- INST is measured by a rule-of-law measure
- TRADE is measured as trade-to-GDP ratio
- MORT and SHARE are instruments

Rodrik example (2)

- Model is defined by

$$\mathrm{INC} = \beta_0 + \beta_1 \mathrm{GEOG} + \beta_2 \mathrm{INST} + \beta_3 \mathrm{TRADE} + \varepsilon,$$

- And moment conditions

$$\mathbb{E}h = \mathbb{E} \begin{pmatrix} \varepsilon \\ \mathrm{GEOG} \cdot \varepsilon \\ \mathrm{MORT} \cdot \varepsilon \\ \mathrm{SHARE} \cdot \varepsilon \end{pmatrix} = 0.$$

- Interest is in

$$\theta = (\beta_0, \beta_1, \beta_2, \beta_3)$$

Rodrik example (3)

All data available:

$$h = \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{pmatrix}, \quad S_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

constant
distance to equator
settler mortality
trade share constructs

Rodrik example (3)

No instruments available

$$h = \begin{pmatrix} h_1 \\ h_2 \\ \times \\ \times \end{pmatrix}, \; S_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

constant
distance to equator
settler mortality
trade share constructs

Rodrik example (3)

Only trade share constructs available

$$h = \begin{pmatrix} h_1 \\ h_2 \\ \times \\ h_4 \end{pmatrix}, \; S_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

constant
distance to equator
settler mortality
trade share constructs

Rodrik example (3)

Dependent or explanatory variable missing

$$h = \begin{pmatrix} \times \\ \times \\ \times \\ \times \end{pmatrix}, \; S_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

constant
distance to equator
settler mortality
trade share constructs

Conditional moments

- Consider the conditional moments

$$\mathbb{E}\left(Rh(X,\theta_0) \mid R \right)$$

- For example, if settler mortality is missing, $R = S_3$,

$$\mathbb{E}\left(\left. \begin{matrix} \varepsilon \\ \mathrm{GEOG} \cdot \varepsilon \\ \color{red}{0} \\ \mathrm{SHARE} \cdot \varepsilon \end{matrix} \right| R = S_3 \right)$$

Conditional moment conditions

Restriction

$$\mathbb{E}\left(Rh(X,\theta_0)\middle|\, R\right) = 0$$

- observable model holds for each subpopulation
- weaker than missing completely at random ($X \perp R$)

Identification

$$\mathbb{E}\left(Rh(X,\theta_0)\middle|\, R\right) = 0 \Leftrightarrow \theta = \theta_0$$

Efficient estimation: Overview

- Assume that θ_0 can be estimated with each subsample (i.e. θ_0 is identifiable in each subpopulation)

1. Estimate θ_0 optimally with each subsample
2. Determine the optimal linear combination of the subsample estimators

- This estimator is asymptotically efficient

Subsample estimators

- Form J subsamples based on the missing data pattern
- G_j is the subsample with pattern S_j
- n_j is the number of observations in subsample G_j
- Denote j-th subsample moment by

$$\bar{h}_{n,j}(\theta) = \frac{1}{n_j} \sum_{i \in G_j} R_i h(X_i, \theta)$$

- Each subsample provides a consistent estimator for θ_0,

$$\hat{\theta}_{j,n} = \operatorname{argmin}_\theta \bar{h}_{n,j}(\theta)' W^*_{j,n} \bar{h}_{n,j}(\theta)$$

Subsample estimators

- Asymptotic variance of j-th estimator is

$$\Lambda_j = \left(D_j' \left(S_j \Omega_j S_j \right)^+ D_j \right)^{-1}$$

- Estimators have different variance matrices
- Note: expected derivatives D_j and variances Ω_j can differ across subpopulations

Matrix weights

- Each subsample estimator is consistent
- Any weighted average is consistent
- Use matrix weights $A_{j,n}$ that sum up to I
- Class of matrix weighted estimators

$$\hat{\theta}_{A(n)} = \sum_{j=1}^{J} A_{j,n}\hat{\theta}_{j,n},$$

  with $\sum_{j=1}^{J} A_{j,n}$
- Look for matrix weights that minimize MSE

Complete and available case

- Naive estimators can be seen as special cases
- Available case: equal weight to each subsample
- Complete case: all weight to complete subsample

Optimal weights

- An optimal linear combination exists
- Optimal estimator lets

$$A^*_{j,n} \to A^*_j = B^{-1}\, p_j \Lambda_j^{-1}$$

where $B = \sum_{k=1}^{J} p_k \Lambda_k^{-1}$

- Intuition: more weight on more precise estimators
- Asymptotic variance of optimal estimator is B^{-1}

Semiparametric efficiency

Definition

The semiparametric efficiency bound is a lower bound on the variance of any regular semiparametric estimator

Theorem

The semiparametric efficiency bound for θ_0 is equal to B^{-1}

Remark

Remark 1

Procedure extends to

- two-step, iterative, and continuous updating GMM estimators
- (generalized) empirical likelihood estimators
- inverse probability weighting estimators

Remark 2

Can include non-identifying subsamples to increase precision, e.g. observations with insufficient instruments

IV: setup

Linear IV

- Linear IV model with one endogenous variable and two instruments, no exogenous variables or constant
- $\mathbb{E}\begin{pmatrix} Z_1(y - X'\beta) \\ Z_2(y - X'\beta) \end{pmatrix} = 0,$

Two similar, normalized, partially missing instruments

- $\text{var}\left(Z\left(y - X'\beta\right)\middle| R\right) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$
- Same correlation with the endogenous variable
- Each instruments is equally likely to be missing
- At least one instrument is observed

IV: results

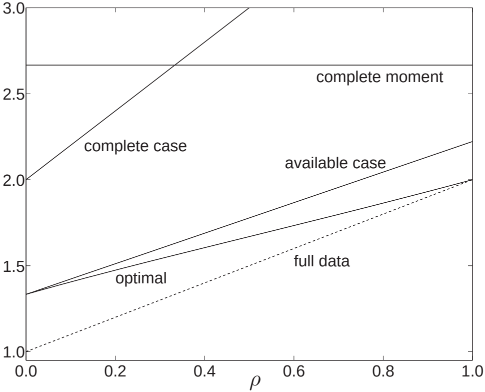


Figure: Asymptotic variance for estimators of β as a function of ρ.

Rodrik

	Rodrik	Available	Optimal
GEOG	-0.70 (0.50)	0.35 (0.16)	0.16 (0.05)
INST	2.00 (0.54)	0.47 (0.34)	0.86 (0.07)
TRADE	-0.30 (0.24)	0.50 (0.40)	0.08 (0.06)

Table: Standard errors in brackets. Sample size for Rodrik: 80. Sample size for available case and optimal: 140.

Arellano and Bond (1991)

- Propose an estimator for a linear dynamic panel data model

$$
\begin{aligned}
& y_{i,t} = \alpha_i + \rho y_{i,t-1} + \varepsilon_{i,t}, \; t = 1, \ldots, T, \\
& \mathbb{E}(\alpha_i) = 0, \\
& \mathbb{E}(\varepsilon_{i,t}) = 0, \\
& \mathbb{E}(\varepsilon_{i,s}\varepsilon_{i,t}) = 0 \text{ if } t \neq s.
\end{aligned}
$$

- Interest in autoregressive parameter ρ.

Arellano and Bond: estimation

Estimation using, e.g. $T = 5$:

Difference	Instruments
$\Delta y_{i,3} = \rho \Delta y_{i,2} + \Delta \varepsilon_{i,3}$	$y_{i,1}$
$\Delta y_{i,4} = \rho \Delta y_{i,3} + \Delta \varepsilon_{i,4}$	$y_{i,1},\ y_{i,2}$
$\Delta y_{i,5} = \rho \Delta y_{i,4} + \Delta \varepsilon_{i,5}$	$y_{i,1},\ y_{i,2},\ y_{i,3}$

Arellano and Bond: missing data patterns

	Missing:			
	None	$y_{i,1}$	$y_{i,4}$	$(y_{i,1}, y_{i,4})$
$y_{i,1}\Delta\varepsilon_{i,3}$	X	·	X	·
$y_{i,1}\Delta\varepsilon_{i,4}$	X	·	·	·
$y_{i,1}\Delta\varepsilon_{i,5}$	X	·	·	·
$y_{i,2}\Delta\varepsilon_{i,4}$	X	X	·	·
$y_{i,2}\Delta\varepsilon_{i,5}$	X	X	·	·
$y_{i,3}\Delta\varepsilon_{i,5}$	X	X	·	·

Table: Available moment conditions are indicated by an X.

Arellano and Bond: Employment equations

- Panel of 140 companies, 7 years
- Employment in company i at time t, is modelled as:

$$n_{i,t} = \gamma_0 + \alpha_1 n_{i,t-1} + \alpha_2 n_{i,t-2} + \beta(L)x_{i,t} + \lambda_t + \eta_i + \nu_{i,t},$$

where $x_{i,t}$ is a vector of exogenous variables, λ_t is a year-specific effect, α_i is a firm-specific effect, and $\nu_{i,t}$ is an error term

Arellano and Bond: data

The panel data set is unbalanced:

Pattern	% companies
XXXXXXXOO	44.29
OXXXXXXXO	27.86
OXXXXXXXX	13.57
XXXXXXXXX	10
XXXXXXXXO	2.86
OOXXXXXXX	1.43
...	

Table: Missing data pattern in Arellano and Bond (1991)

Results

	Available case	Optimal	Δ %
$n_{i,t-1}$	0.71 (0.15)	0.60 (0.06)	61
$w_{i,t}$	-0.61 (0.07)	-0.55 (0.05)	19
$w_{i,t-1}$	0.41 (0.11)	0.28 (0.06)	44
$k_{i,t}$	0.36 (0.04)	0.38 (0.03)	19
$k_{i,t-1}$	-0.06 (0.06)	-0.02 (0.03)	39
$ys_{i,t}$	0.63 (0.13)	0.45 (0.13)	5
$ys_{i,t-1}$	-0.72 (0.18)	-0.71 (0.14)	23

w is real product wage, k is log gross capital, ys is log of industry output. Year dummies included.

What can be relaxed

1. GMM
2. Subsample identification
3. Random sample
4. No selection

Random sample

- Independent observations implies independent subsamples
- Estimation is done using the moment conditions

$$\mathbb{E}(\tilde{h}) = \mathbb{E}\begin{pmatrix} h | R = S_1 \\ h | R = S_2 \end{pmatrix}$$

- Optimal weights come from $\mathrm{var}(\tilde{h}) \sim \begin{pmatrix} \Omega_1 & 0 \\ 0 & \Omega_2 \end{pmatrix}$
- Generalization requires estimator of off-diagonal blocks
- Bootstrap the subsample estimators

Missing at random

- Assume that there exists covariates V that remove the dependence between h and R at θ_0:

$$\mathbb{E}\left(h(X,\theta_0)|\, R = S_j, V\right) = \mathbb{E}\left(h(X,\theta_0)|\, V\right)$$

- This resembles unconfoundedness
- Let $p_j(V) = P(R = S_j|V)$ be the probability of seeing missing data pattern j conditional on V
- This is a mean-independence version of missing at random (MAR), where $X \perp R|V$

Inverse probability weighting

- IPW moment condition, for each j, is

$$\tilde{h}_j(X,V,R,\theta) = \frac{1\{R = S_j\}}{p_j(V)} Rh(X,\theta)$$

- Using the LIE, $\mathbb{E}(\tilde{h}_j(X,V,R,\theta)) = 0$
- Set of moment conditions to use: $\left\{\tilde{h}_j,\ j = 1,\cdots,J\right\}$
- Combine subsample estimators optimally, as before

Conclusion

- missing data is a very common problem
- often, incomplete observations are informative
- we show how to efficiently combine all information
- the estimator is easy to implement
- it extends to e.g. inverse probability weighting, GEL

Adjusted criterion function

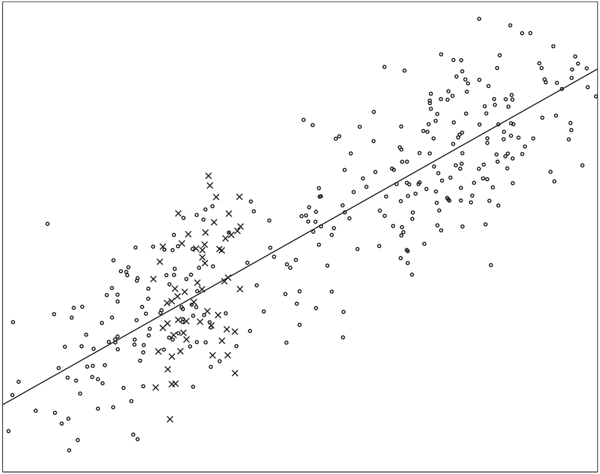
- Full data optimal GMM estimator minimizes $\left(1/n \sum_{i=1}^{n} h(X_i, \theta)\right)' W_n \left(1/n \sum_{i=1}^{n} h(X_i, \theta)\right)$, $W_n \to \Omega^{-1}, \ \Omega = \mathrm{var}(h(X, \theta_0))$.
- $G_j = \{i : R_i = S_j\}$ is the group of observations with pattern j
- $n_j = \#\{G_j\}$ the number of observations in each group.
- Adjusted criterion function $\sum_{j=1}^{J} \left(1/n_j \sum_{i \in G_j} R_i h(X_i, \theta)\right)' W_{j,n} \left(1/n_j \sum_{i \in G_j} R_i h(X_i, \theta)\right)$, $W_{j,n}$ are group-specific weight matrices.

Optimal GMM with missing data

- Optimal GMM estimator for missing data: $\hat{\theta}_n^*$ by letting $W_{j,n} \to W_j^* = p_j(S_j \Omega_j S_j)^+$.
- Limiting distribution: $\sqrt{n}(\hat{\theta}_n^* - \theta_0) \xrightarrow{d} N\left(0, B^{-1}\right)$.
- Asymptotically efficient.
- Interpretation: efficient GMM estimator using all Jq moment conditions $\mathbb{E}(h(X, \theta_0) | R = S_j) = 0$.

MCAR versus mean-independence

y_i



x_i

Attrition (1)

- Toy example.
- Health club: new members are weighted at registration.
- A random sample is selected for a reweighting after 3 months.
- We are interested in the weight change in new members.

Attrition (2)

- For a person's weight $X_{i,t},\ t = 1, 2$, assume $X_{i,t} = \alpha_i + \mu_t + \varepsilon_{i,t}$.
- Assume $\mathbb{E}(\alpha_i) = \mathbb{E}(\varepsilon_{i,t}) = 0$ and $\mathbb{E}(\varepsilon_{i,1}\varepsilon_{i,2}) = 0$.
- Normalize $\text{var}(\alpha_i) + \text{var}(\varepsilon_{i,t}) = 1$, $\rho = \text{var}(\alpha_i)/(\text{var}(\alpha_i) + \text{var}(\varepsilon_{i,t}))$
- $J = 2,\ S_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ S_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$
- $\text{var}\left(\left(X_{i,1},\ X_{i,2}\right) | R_i\right) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$
- For illustration, I will assume ρ known.
- $p_1 = 0.5$.

Attrition (3)

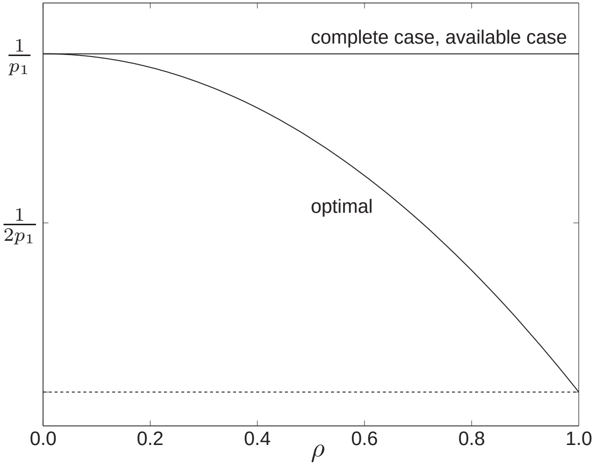


Figure: Asymptotic variances of $\hat{\mu}_2$ as a function of ρ.

Attrition (4)

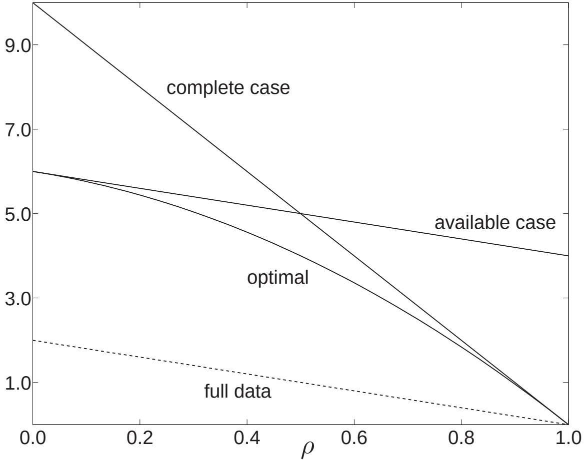


Figure: Asymptotic variances of $\hat{\mu}_2 - \hat{\mu}_1$ as a function of ρ.

Dynamic panel data (1)

$$\begin{aligned}
&y_{i,t} = \alpha_i + \rho y_{i,t-1} + \varepsilon_{i,t},\ t = 1, \ldots, T, \\
&\mathbb{E}(\alpha_i) = 0, \\
&\mathbb{E}(\varepsilon_{i,t}) = 0, \\
&\mathbb{E}(\varepsilon_{i,s}\varepsilon_{i,t}) = 0 \text{ if } t \neq s.
\end{aligned}$$

Interest in autoregressive parameter ρ.

Dynamic panel data (2)

Estimation using Arellano and Bond (1991). $T = 5$:

$\Delta y_{i,3} = \rho \Delta y_{i,2} + \Delta \varepsilon_{i,3}$	$y_{i,1}$
$\Delta y_{i,4} = \rho \Delta y_{i,3} + \Delta \varepsilon_{i,4}$	$y_{i,1},\ y_{i,2}$
$\Delta y_{i,5} = \rho \Delta y_{i,4} + \Delta \varepsilon_{i,5}$	$y_{i,1},\ y_{i,2},\ y_{i,3}$

Dynamic panel data (3)

	Missing components			
	None	$y_{i,1}$	$y_{i,4}$	$(y_{i,1}, y_{i,4})$
$y_{i,1}\Delta\varepsilon_{i,3}$	X	·	X	·
$y_{i,1}\Delta\varepsilon_{i,4}$	X	·	·	·
$y_{i,1}\Delta\varepsilon_{i,5}$	X	·	·	·
$y_{i,2}\Delta\varepsilon_{i,4}$	X	X	·	·
$y_{i,2}\Delta\varepsilon_{i,5}$	X	X	·	·
$y_{i,3}\Delta\varepsilon_{i,5}$	X	X	·	·

Dynamic panel data (4)

- $T = 9$ as in Blundell and Bond (1998),
- 10 patterns: at most one $y_{i,t}$ missing per i,
- $p = 0.02,\ 0.06$, (82%, 46% complete)
- $\text{var}(e) = 1$,
- $\Omega_j = \Omega$,
- Continuous updating GMM,
- $n = 10000,\ S = 1000$.

Dynamic panel data (5)

var(α)	ρ	p	cc	ac	opt	full
0.1	0.2	0.02	1.29	1.23	1.18	1
		0.06	2.37	1.34	1.27	1
	0.5	0.02	1.82	1.77	1.69	1
		0.06	3.35	2.50	2.25	1
1	0.2	0.02	1.91	1.70	1.68	1
		0.06	3.75	2.35	2.21	1
	0.5	0.02	5.10	4.75	4.59	1
		0.06	8.61	5.85	5.33	1

Table: Relative variance of the complete case (cc), available case (ac) and optimal (opt) estimator with respect to full data estimator.