Missing end🚫genous variables
in conditional-moment-restriction models

Antonio Cosma[*]
Andreï V. Kostyrka[†]
Gautam Tripathi[†]

[*] UNIVERSITÀ DEGLI STUDI DI BERGAMO | Department of Management

[†] uni.lu | UNIVERSITÉ DU LUXEMBOURG
Department of Economics and Management (DEM)

Internal Research Seminars in Economics and Management
University of Luxembourg
24th of October 2023

Presentation structure

1. Motivation and empirical application

2. Identification in the presence of missing observations

3. Efficient estimation and inference under missingness at random

4. Simulation study

Motivation and empirical application

Outline and motivation

We extend the existing literature on missing-data models defined by conditional moment restrictions by:

1. Computing the efficiency bound and proposing a 'doubly robust' estimator that attains the efficiency bound
2. Explicitly addressing the role of non-missing endogenous variables in obtaining efficiency gains by using the entire sample
3. Carrying out simulations showing that the efficiency gains from using the proposed estimator are comparable with the maximum gains DGP can deliver, and applying it to a real model and data

Literature

- Robins et al. (1994, *JASA*) propose a new class of consistent semi-parametric estimators when the data are missing at random
- Chen, Hong, Tarozzi (2008, **AoS**) derive semi-parametric efficiency bounds for missing-data models defined by **un**conditional moment restrictions
- Graham (2011, *Ecta*) introduces the **equivalence** result for unconditional moment restrictions
- Hristache and Patilea (2016, *ET*; 2017, *Biometrika*) extend the equivalence result to conditional moment restrictions

Conditional-moment-restriction models

The models of interest are defined by a *conditional* moment restriction:

$$\exists \theta^* : \mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X] = 0$$

- Earning equation (Mincer)

$$\log \mathit{wage}^* = \alpha + \gamma\, \mathit{educ} + U$$

$$\mathbb{E}[\log \mathit{wage}^* - \alpha - \gamma\, \mathit{educ} \mid \mathit{parent\_educ}] = 0$$

- *CEO* succession in family firms (Bennedsen et al., 2007)

$$\mathit{perf} = \alpha + \gamma\, \mathit{fam\_succ}^* + \beta' X_{\text{industry}} + U$$

$$\mathbb{E}[\mathit{perf} - \alpha - \beta' X_{\text{industry}} - \gamma\, \mathit{fam\_succ}^* \mid \mathit{boy\_1}^{st}, X_{\text{industry}}] = 0$$

Missing data

In the **target population** model,

$$\mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X] = 0$$

- Y^* contains endogenous variables (outcome or explanatory) that are not observed for some units
- Z and X are vectors of always observed endogenous and exogenous variables respectively

At the *observational* level (in the **realised sub-population**):

$$Y := DY^* + (1 - D)\mathbf{m},$$

where $D = 1$ if all coordinates of Y^* are observed (0 otherwise) and **m** is a symbol for missing values (e. g. '–999', 'NA', '.' in packages and survey codebooks).

Estimation objective

Z	X	D	Y	Y^*
2.66	0.34	1	3.50	3.50
1.94	0.37	1	2.12	2.12
1.05	0.38	0	**m**	2.09
−0.98	0.38	1	−0.52	−0.52
0.91	0.38	0	**m**	1.37
1.92	0.39	0	**m**	3.24
4.15	0.50	1	5.06	5.06
1.42	0.57	1	2.36	2.36
2.39	0.63	1	3.35	3.35
1.59	0.65	0	**m**	2.53
−0.48	0.66	1	−0.28	−0.28
1.18	0.69	0	**m**	1.70

- Using observations $(Y_i, D_i, X_i, Z_i)_{i=1}^n$, efficiently estimate θ^*
- A data set example can be seen on the right
- Call the subsample with $D = 1$ the **validation sample**

Estimation objective

Z	X	D	Y	Y^*
2.66	0.34	1	3.50	3.50
1.94	0.37	1	2.12	2.12
1.05	0.38	0	**m**	2.09
−0.98	0.38	1	−0.52	−0.52
0.91	0.38	0	**m**	1.37
1.92	0.39	0	**m**	3.24
4.15	0.50	1	5.06	5.06
1.42	0.57	1	2.36	2.36
2.39	0.63	1	3.35	3.35
1.59	0.65	0	**m**	2.53
−0.48	0.66	1	−0.28	−0.28
1.18	0.69	0	**m**	1.70

- Using observations $(Y_i, D_i, X_i, Z_i)_{i=1}^{n}$, efficiently estimate θ^*
- A data set example can be seen on the right
- Call the subsample with $D = 1$ the **validation sample**

Practical application

Angrist & Evans (1998, AER): female labour supply model.

$$Y^* = \alpha + X_{\text{incl}}'\beta + \gamma \cdot \mathit{MOREKIDS} + U, \quad \mathbb{E}(U \mid X_{\text{incl}}, X_{\text{excl}}) = 0$$

- Y^*: labour income of a working mother
- Endogenous: $\mathit{MOREKIDS} := \mathbb{I}(\text{mother has} \geq 3 \text{ kids})$
- X_{incl}: age, age at first birth, sex of the first child
- X_{excl}: $\mathbb{I}(\text{two boys})$, $\mathbb{I}(\text{two girls})$

Data: 1990 PUMS sub-sample of $n = 260\,286$ white females.

Question: What if income is not always reported?

Counterfactual exercise

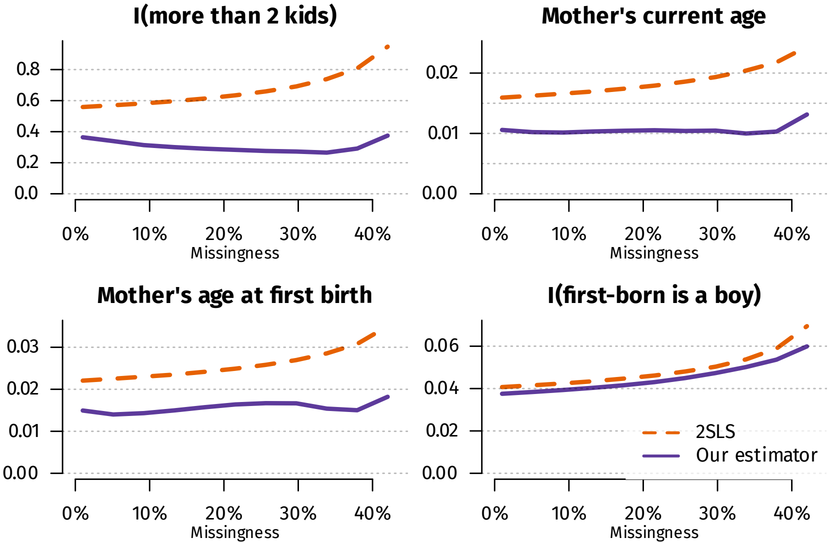
1. We induce missingness of various strength (1–40%) in labour income (Y^*)
2. We estimate a model of female labour supply using 2 methods: 2SLS (original) and the proposed method

- How robust is the original model to the presence of missing observations?
- Can our methodology be used to address the issue of missingness in a useful manner?

Practical findings

- 260 000 observations with 40% missing outcomes may be not enough to find a significant effect of having more than 2 children on female labour income
  - Case-wise deletion (default option in software) leaves 150 000 observations
  - At the 10%, 5%, 1% level, the coefficient on *MOREKIDS* seems insignificant in **62%, 70%, and 99%** simulations respectively
- Our method produces more reliable and accurate estimates
  - At the 10%, 5%, 1% level, the coefficient on *MOREKIDS* seems insignificant in **1%, 22%, and 30%** simulations respectively
  - Findings are stable for a wide range of internal tweaking parameters (bandwidths, kernel, smoother degree)

Standard errors in the practical application



Missingness mechanism

Assume the **missing-at-random** (MAR) mechanism:

$$D \perp\!\!\!\perp Y^* \mid X, Z$$

It is a form of **selection on observables**:

$$\mathbb{P}(D = 1 \mid Y^*, X, Z) = \mathbb{P}(D = 1 \mid X, Z)$$

The probability of retention in the sample only depends on the observable *X* and *Z*.

Example: IV reg. with missing outcomes

Let $X = (X_{\text{incl}}, X_{\text{excl}})$.

Target model: $Y^* = \alpha^* + X_{\text{incl}}'\beta^* + Z'\gamma^* + U, \quad \mathbb{E}[U \mid X] = 0.$

Selection model: $D = \pi(Z, X) + V, \quad \mathbb{E}[V \mid Z, X] = 0.$

$$\begin{aligned}
&D \perp\!\!\!\perp Y^* \mid Z, X \\
&\quad \iff [\pi(Z,X) + V] \perp\!\!\!\perp [\alpha^* + X_{\text{incl}}'\beta^* + Z'\gamma^* + U] \mid Z, X \\
&\quad \iff V \perp\!\!\!\perp U \mid Z, X
\end{aligned}$$

MAR properties

- One do not need to specify the selection equation
  - As opposed to Heckman-like or selection-on-**un**observables missingness mechanisms
- Consistent with the semi-parametric framework
  - No parametric assumptions about the distribution of the innovations *U* required
- One can use the entire sample to increase the efficiency of the estimator
- In the observation equation, one may include endogenous variables where the source of endogeneity is different from the selection process

Drawback: not testable.

Identification in the presence of missing observations

Equivalence result: Graham ’11, H&P ’16

Under MAR, the model in the target population

$$\mathbb{E}[g(Y^*, Z, X, \theta^*) \,|\, X] = 0, \quad D \perp\!\!\!\perp Y^* \mid X, Z$$

is **equivalent** to the one in the realised sub-population

$$\begin{cases} \mathbb{E}\Big[\dfrac{D}{\pi(X,Z)} g(Y, Z, X, \theta^*) \,\Big|\, X\Big] = 0 & (1) \\ \mathbb{E}\Big[\dfrac{D}{\pi(X,Z)} - 1 \,\Big|\, Z, X\Big] = 0 & (2) \end{cases}$$

(2) defines the **propensity score** $\pi(X, Z) := \mathbb{E}(D \mid X, Z)$.

- $\pi(X, Z)$ is estimated non-parametrically, even if it is parametrically specified or it is fully known
- The larger conditioning set in (2) enables efficiency gains

Efficiency bound

Following Ai & Chen (2003, 2012), we provide the efficiency bound implied by the model $\mathbb{E}[\rho(Y, Z, X, \theta^*) \mid X] = 0$.

$$\rho(Y, Z, X, \theta^*) := \frac{D}{\pi(X, Z)} g(Y, Z, X, \theta^*) + \left(1 - \frac{D}{\pi(X, Z)}\right) \mu(X, Z, \theta^*)$$

$\mu(X, Z, \theta^*) := \mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X, Z]$ is the **non-parametric imputation** of g (under MAR, $\mu = \mathbb{E}[g(Y, Z, X, \theta^*) \mid X, Z, D = 1]$.

$$\text{l.b.}(\theta^*) := (\mathbb{E} J' \Omega_\rho^{-1} J)^{-1}$$

$J := \frac{\partial}{\partial \theta^*} \mathbb{E}[\rho \mid X]$ = Jacobian of the moment function,
$\Omega_\rho := \mathbb{E}\left[\frac{Dgg'}{\pi^2} \;\middle|\; X\right] - \frac{1-\pi}{\pi} \mathbb{E}[\mu\mu' \mid X]$ = variance of the mom. fun.

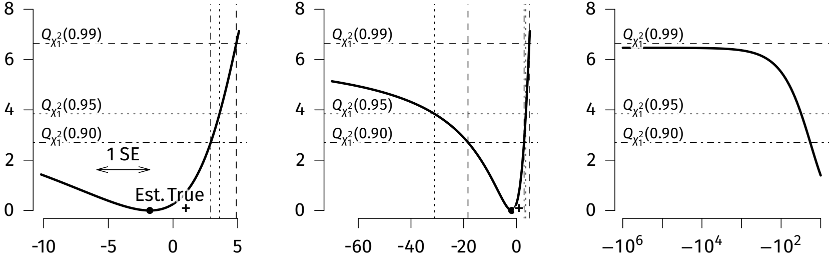
Results so far

1. **According to the selection mechanism, estimation only on the validation sample may not deliver consistent estimates**

Why efficiency is important

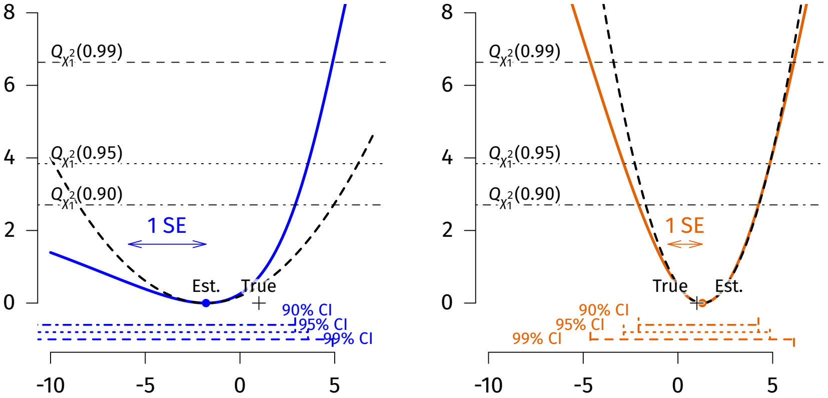
If only the validation sample is used, the confidence intervals (even although asymptotically valid) may be unbounded in finite samples.

Example: simple linear regression, 1 endogenous variable, 1 instrument, n = 500, 36% missingness rate. Confidence interval for the slope:



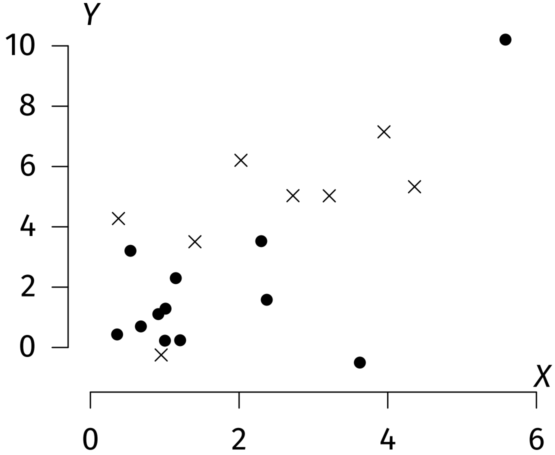
Efficiency and confidence intervals

ELR statistic for the slope value: VS only, full-sample.
Dashed: Wald statistic for the slope value.



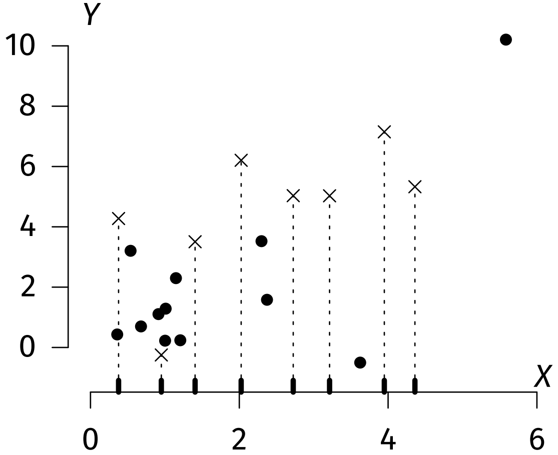
Imputation without endogenous variables

$$Y = \beta_0 + \beta_1 X + U, \quad \mathbb{E}(U \mid X) = 0$$



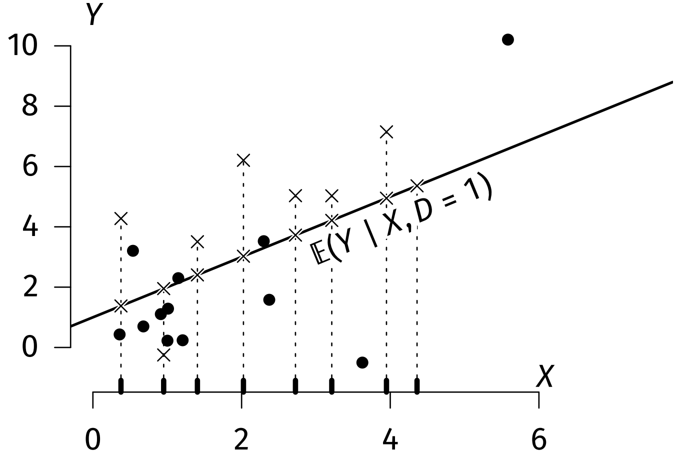
Imputation without endogenous variables

$$Y = \beta_0 + \beta_1 X + U, \quad \mathbb{E}(U \mid X) = 0$$



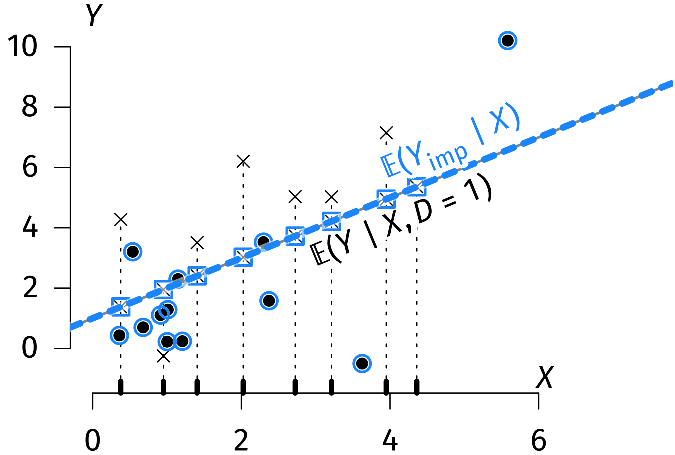
Imputation without endogenous variables

$$Y = \beta_0 + \beta_1 X + U, \quad \mathbb{E}(U \mid X) = 0$$



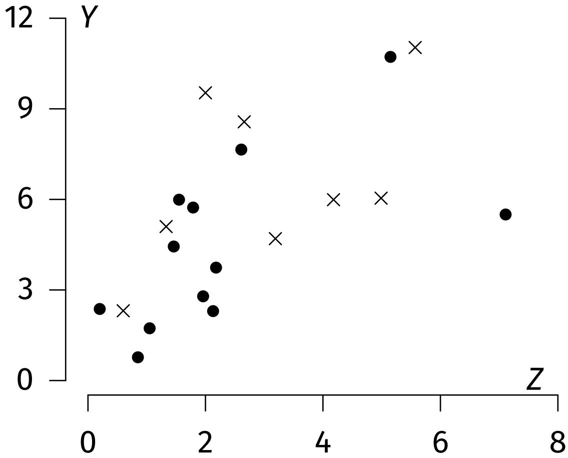
Imputation without endogenous variables

$Y = \beta_0 + \beta_1 X + U, \quad \mathbb{E}(U \mid X) = 0, \quad Y_{\text{imp}} = DY + (1 - D)\mathbb{E}(Y \mid X, D = 1)$



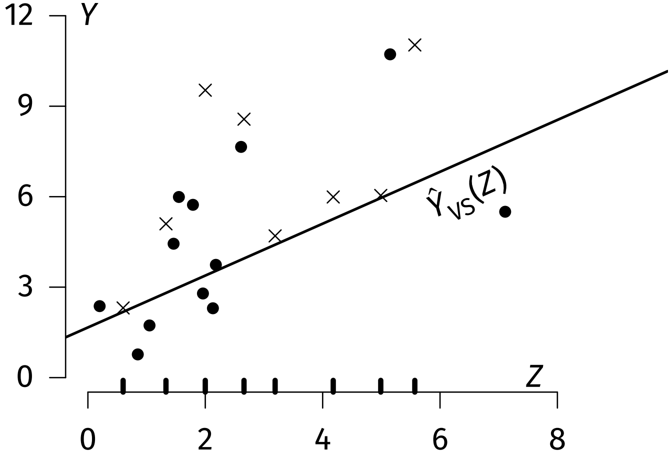
Imputation with endogenous variables

$$Y = \gamma_0 + \gamma_1 Z + U, \quad Z = \zeta_0 + \zeta_1 X + V, \quad \mathbb{E}(U \mid X) = 0, \quad \mathbb{E}(U \mid X, Z) \neq 0$$



Imputation with endogenous variables

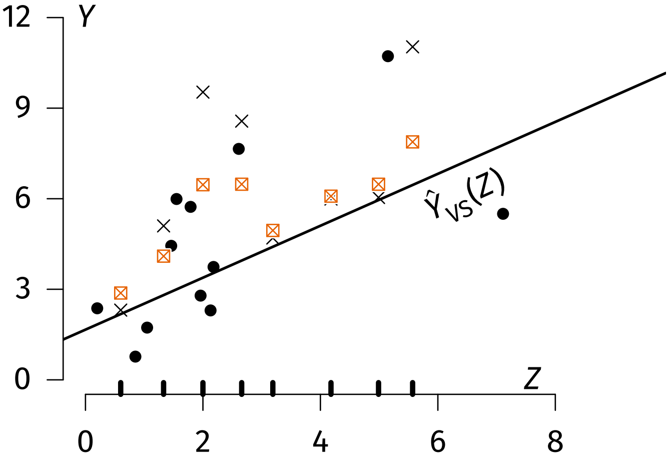
$$Y = \gamma_0 + \gamma_1 Z + U, \quad Z = \zeta_0 + \zeta_1 X + V, \quad \mathbb{E}(U \mid X) = 0, \quad \mathbb{E}(U \mid X, Z) \neq 0$$



Imputation with endogenous variables

$$Y = \gamma_0 + \gamma_1 Z + U, \quad Z = \zeta_0 + \zeta_1 X + V, \quad \mathbb{E}(U \mid X) = 0, \quad \mathbb{E}(U \mid X, Z) \neq 0$$

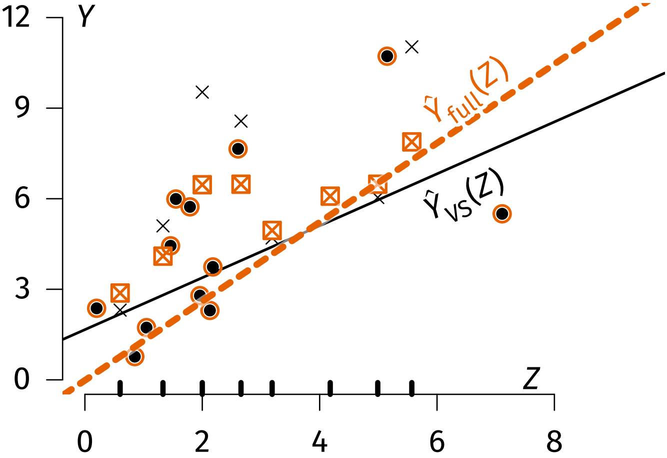
$$Y_{\text{imp}} = DY + (1 - D)\mathbb{E}(Y \mid X, Z, D = 1)$$



Imputation with endogenous variables

$$Y = \gamma_0 + \gamma_1 Z + U, \quad Z = \zeta_0 + \zeta_1 X + V, \quad \mathbb{E}(U \mid X) = 0, \quad \mathbb{E}(U \mid X, Z) \neq 0$$

$$Y_{\text{imp}} = DY + (1 - D)\mathbb{E}(Y \mid X, Z, D = 1)$$



Results so far

1. According to the selection mechanism, estimation only on the validation sample may not deliver consistent estimates
2. **To have efficiency gains, one needs a larger conditioning set for the propensity score than for the conditional moment restriction of the model**

ρ for efficient estimation of a linear model

$$Y^* = \alpha_0 + \gamma_0 Z + U, \quad \mathbb{E}(U \mid X) = 0 \qquad \text{(MAR)}$$

$$\begin{cases} \mathbb{E}\Big[\frac{D(Y-\alpha_0-\gamma_0 Z)}{\pi(X,Z)} \;\Big|\; X\Big] = 0 \\ \mathbb{E}\Big[\frac{D}{\pi(X,Z)} - 1 \;\Big|\; Z, X\Big] = 0 \end{cases}$$

$$\mu(X, Z, \alpha_0, \gamma_0) = \mathbb{E}[Y - \alpha_0 - \gamma_0 Z \mid X, Z]$$

$$\rho = \frac{D}{\pi(X,Z)}\big(Y - \alpha_0 - \gamma_0 Z\big) + \Big[1 - \frac{D}{\pi(X,Z)}\Big]\mathbb{E}[Y - \beta_0 - \gamma_0 Z \mid X, Z]$$

$\pi(X, Z)$ and $\mu(X, Z, \alpha_0, \gamma_0)$ are estimated non-parametrically via kernel methods (Nadaraya–Watson, LOESS).

Double robustness against misspecification

$$\rho = g + \left(\frac{D}{\pi} - 1\right)(g - \mu)$$

π and μ are unknown functions that must be estimated non-parametrically ($\hat\pi$, $\hat\mu$).

However, a researcher may use easier parametric estimators $\tilde\pi$ and $\tilde\mu$ instead of $\hat\pi$ and $\hat\mu$, with misspecification error

$$\text{'noise'} = \left(\frac{\pi}{\tilde\pi} - 1\right)(\tilde\mu - \mu)$$

Double robustness: θ^* is consistently estimable when *either* the selection model for D (defining $\pi(Z, X)$) *or* the model for imputing g (defining $\mu(X, Z, \theta^*)$) is correctly specified (because then, $\mathbb{E}(\text{'noise'} \mid Z, X) = 0$).

Results so far

1. According to the selection mechanism, estimation only on the validation sample may not deliver consistent estimates
2. To have efficiency gains, one needs a larger conditioning set for the propensity score than for the conditional moment restriction of the model
3. **The efficient moment condition uses non-parametric imputation for observations even without missingness**

Efficient estimation and inference under missingness at random

Smooth empirical likelihood (SEL)

- (S)EL is a non-parametric method for testing and estimating (Owen, 1988; Kitamura, Tripathi, Ahn, 2004)
- EL estimators based on unconditional moment restrictions are equivalent to optimally weighted GMM estimators
- SEL imposes a conditional moment restriction on the parameters and data
- Parametric restrictions can be tested using a nonparametric version of Wilk's theorem (Qin & Lawless, 1994)

SEL problem

Take any **conditional**-moment-restriction model

$$\mathbb{E}[h(A, \theta) \mid X] = 0$$

(A, X): generic vector of data, $\{(A_i, X_i)\}_{i=1}^n$: a random sample.

Given kernel weights $w_{ij} := \frac{K(X_i - X_j)}{\sum_{k=1}^n K(X_i - X_k)}$, find the optimal discrete distribution $\{p_{ij}\}$ to enforce the sample analogue of the restriction above $(h_j(\theta) := h(A_j, \theta))$:

$$\max_{p_{ij}>0} \sum_{i=1}^n \sum_{j=1}^n w_{ij} \log p_{ij} \quad \text{s.t.} \quad \begin{cases} \sum_{j=1}^n p_{ij} = 1 & \forall i = \overline{1, n} \\ \sum_{j=1}^n p_{ij} h_j(\theta) = 0 & \forall i = \overline{1, n} \end{cases}$$

Estimation with SEL

For any θ, the solution to the optimisation problem above is

$$\hat{p}_{ij}(\theta) = \frac{w_{ij}}{1 + \hat{\lambda}_i'(\theta) h_j(\theta)},$$

where $\hat{\lambda}_i$ are the Lagrange multipliers for $\sum_{j=1}^n p_{ij} h_j(\theta) = 0$.

Define the value function $\mathrm{SEL}(\theta) := \sum_{i=1} [\sum_{j=1}^n w_{ij} \log \hat{p}_{ij}(\theta)]$.

SEL estimation algorithm:

1. Find $\hat{\lambda}_i(\theta)$ (n concave 1-dimensional problems)
2. Compute $\hat{p}_{ij}(\theta)$ and $\mathrm{SEL}(\theta)$
3. Find $\hat{\theta} := \arg\max_\theta \mathrm{SEL}(\theta)$

Inference with SEL

- Asymptotic normality: $\sqrt{n}(\hat\theta - \theta^*) \sim \mathcal{N}(0, V)$
- Asymptotic efficiency: $V = \mathrm{l.b.}(\theta^*)$
- Hypothesis testing: for a parametric restriction $\mathscr{H}_0\colon R(\theta) = 0$, find $\hat\theta_R := \arg\max_{\theta\colon R(\theta)=0} \mathrm{SEL}(\theta)$. Then, the SELR statistic $2[\mathrm{SEL}(\hat\theta) - \mathrm{SEL}(\hat\theta_R)]$ is asymptotically $\chi^2_{\dim R}$
- α% confidence regions: invert the SELR statistic, find $\{\theta\colon \mathrm{SELR}(\theta) \le Q_{\chi^2_{\dim\theta^*}}(\alpha\%)\}$
- Standard errors: $\mathrm{SE}(\hat\theta^{(j)}) = \sqrt{\mathrm{diag}[-\nabla^2_{\mathrm{SEL}}(\hat\theta)]^{-1}}$

SEL and sieve minimum distance (SMD)

Ai & Chen (2003) proposed the SMD estimator:

$$m(\theta) := \mathbb{E}[\rho(\theta) \mid X], \qquad \Omega(\theta) := \mathrm{Var}[\rho(\theta) \mid X]$$

Using $\hat{m}_i(\theta) := \sum_{j=1}^n w_{ij}\rho_j(\theta)$ and $\hat{\Omega}_i(\theta) := \sum_{j=1}^n w_{ij}\rho_j(\theta)\rho_j'(\theta)$, minimise the continuous updating objective:

$$\tilde{\theta} := \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \hat{m}_i'(\theta)\hat{\Omega}_i^{-1}(\theta)\hat{m}_i(\theta)$$

In large samples, $\hat{\theta} \approx \tilde{\theta}$, but

- SEL $\hat{\theta}$ skips the difficult variance estimation ('explicit studentisation')
- SELR is internally studentised and implicitly pivotal

Visualisation of OLS, SMD, and SEL

[Animation.]

Roadmap for efficient estimation

1. Assume missingness at random for a conditional-moment-restriction problem
2. Re-write the problem in (1) as two sequential moment conditions on the observed values
3. Orthogonalise the two functions appearing in the moments of (2), obtain a single CMR based on ρ
4. Using SEL on $\mathbb{E}[\rho(\theta) \mid X] = 0$, obtain efficient estimates of θ

Simulation study

Simulation results (design 1, discrete)

Missingness only in Y^*, one discrete endogenous variable Z, one discrete excluded instrument X, exogenous selection.

$$Y^* = 1 + 1 \cdot Z + U\sigma(X), \ \mathbb{E}(U \mid X) = 0, \qquad \sigma^2(X) = \begin{cases} 16, & X = 0, \\ 1, & X = 1, \end{cases}$$

$$X \sim \text{Bernoulli}(0.6), \quad Z = \mathbb{I}(X + V > 0), \ \pi(X) = \begin{cases} 0.25, & X = 0, \\ 0.9, & X = 1, \end{cases}$$

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right] \quad \Rightarrow \quad \begin{matrix} \text{strong endogeneity} \\ \text{(OLS slope bias 179\%)} \end{matrix}$$

Missingness rate: 36%, max. gains: 31%.

Monte-Carlo simulation summary (design 1)

10 000 simulations for each n. **Theoretical gains: 31%.**
Slope estimator statistics are reported.
VS: validation sample only, FS: full-sample efficient.

n	Est.	Med. bias	Mean bias	SD	$\frac{\text{Var(VS)}}{\text{Var(FS)}}$	$\frac{\text{MSE(VS)}}{\text{MSE(FS)}}$
500	FS	0.042	−0.032	2.02	6.25	6.33
	VS	−0.025	−0.604	5.05		
1000	FS	0.027	−0.027	1.40	1.67	1.69
	VS	0.004	−0.229	1.80		
2000	FS	0.041	0.019	0.96	1.49	1.49
	VS	0.015	−0.081	1.18		
4000	FS	0.034	0.014	0.67	1.39	1.39
	VS	0.022	−0.036	0.79		

Confidence intervals for the slope (design 1)

		Coverage probability			
n	Estimator	Nominal	Empirical	Med. length	% bounded
500	FS	.900	.905	6.66	100
		.950	.952	8.17	100
		.990	.991	11.51	100.0
	VS	.900	.897	8.43	96.9
		.950	.949	10.77	94.1
		.990	.990	16.67	84.2
4000	FS	.900	.904	2.24	100
		.950	.957	2.68	100
		.990	.991	3.55	100
	VS	.900	.903	2.59	100
		.950	.948	3.11	100
		.990	.991	4.18	100

FS = full-sample (efficient), VS = validation-sample only ($D = 1$).

Simulation results (design 2, continuous)

Missingness only in Y^*, one endogenous variable Z, one excluded instrument X.

$$\begin{aligned} Y^* &= 1 + 1 \cdot Z + U\sigma(X), & \mathbb{E}(U \mid X) &= 0, \\ X &\sim \mathrm{Uniform}[0, 1], & Z &= 1 + X + V, \\ \pi(X) &= 0.25 + 0.7\Phi\Bigl(\tfrac{X-0.1}{0.5}\Bigr), & \sigma^2(X) &= 1/15 + (X + 1/3)^2, \end{aligned}$$

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right] \quad \Rightarrow \quad \begin{matrix} \text{strong endogeneity} \\ \text{(OLS slope bias 42\%)} \end{matrix}$$

$\Phi(\cdot)$ is the standard normal CDF (sigmoid).

Missingness rate: 58%, max. gains: 42%.

Monte-Carlo simulation summary (design 2)

10 000 simulations for each n. **Theoretical gains: 42%.**
Slope estimator statistics are reported.
VS: validation sample only, FS: full sample, efficient.

n	Est.	Med. bias	Mean bias	SD	$\frac{\text{Var(VS)}}{\text{Var(FS)}}$	$\frac{\text{MSE(VS)}}{\text{MSE(FS)}}$
500	FS	0.087	0.079	0.147	0.88	1.01
	VS	0.104	0.097	0.138		
1000	FS	0.020	0.010	0.127	1.14	1.14
	VS	0.025	0.011	0.136		
2000	FS	−0.002	−0.009	0.097	1.43	1.44
	VS	−0.001	−0.017	0.116		
4000	FS	−0.000	−0.004	0.064	1.44	1.45
	VS	−0.001	−0.009	0.077		

Confidence intervals for the slope (design 2)

		Coverage probability			
n	Estimator	Nominal	Empirical	Med. length	% bounded
500	FS	.900 .950 .990	.916 .968 .995	.646 .828 1.354	100 100 99.6
	VS	.900 .950 .990	.920 .970 .995	.698 .943 1.820	100 100.0 93.9
4000	FS	.900 .950 .990	.913 .957 .993	.219 .264 .353	100 100 100
	FS	.900 .950 .990	.912 .955 .994	.258 .313 .429	100 100 100

FS = full-sample (efficient), VS = validation-sample only ($D = 1$).

Further developments

- Show that the first-step estimators do not affect the asymptotic variance of the estimator (Ai & Chen 2003, sieve-estimator approach)
- An approach to handle missing *exogenous* regressors
  - Much harder problem due to X being in the structural-model conditioning set
- Develop an optimal bandwidth-selection rule or a convenient rule of thumb for (1) SEL weights, (2) $\hat{\mu}$, (3) $\hat{\pi}$
- Upload an R package for SEL estimation with missing data to CRAN, implement parallel capabilities for faster optimisation
  - A package for quick and accurate gradients is in the making

Conclusions

- We show that if at least one of endogenous variables in the model contains missing values + not all endogenous variables are missing, then, there are efficiency gains compared to the classical complete-case approach
- We derive the efficiency bound and propose an estimator that attains it
- We test its performance in practice and find that it yields empirical gains close to theoretically expected ones
- We apply the estimator to a real data set with introduced missingness and find that confidence intervals based on the efficient estimator are tighter than those of 2SLS

Thank you for your attention!

Now it is time for your questions.

Missingness mechanism – I

Let $X = (X_{\text{incl}}, X_{\text{excl}})$,

$$Y^* = \alpha^* + X_{\text{incl}}'\beta^* + Z'\gamma^* + U \qquad \text{(Target model)}$$

$$D = \pi(Z, X) + V, \qquad \text{(Selection model)}$$

where $\mathbb{E}[U \mid X] = 0$, and $\mathbb{E}[V \mid Z, X] = 0$, so that

$$\begin{aligned} D \perp\!\!\!\perp Y^* \mid Z, X &\Leftrightarrow [\pi(Z, X) + V] \perp\!\!\!\perp [\alpha^* + X_{\text{incl}}'\beta^* + Z'\gamma^* + U] \mid Z, X \\ &\Leftrightarrow V \perp\!\!\!\perp U \mid Z, X. \end{aligned}$$

Missing mechanism – II

This is different from *selection on unobservables*, where the $\mathbb{P}(D = 1 \mid Y^*, X, W) \neq \mathbb{P}(D = 1 \mid X, W)$, as in the following example with both X and W exogenous:

$$\begin{aligned} Y^* &= \alpha + \beta X + U, \\ \widetilde{D} &= \delta_0 + \delta_1 X + \delta_2 W + V, \quad V \not\perp\!\!\!\perp U \mid X, W, \\ D &= \mathbb{I}(\widetilde{D} \geq 0), \end{aligned}$$

Then, even conditioning on (X, W, Y^*, D) would be dependent through the joint distribution of (U, V), as in the Heckman selection model:

$$\mathbb{E}(Y \mid X, W, D = 1) = \alpha + \beta X + \mathbb{E}(U \mid V > -\delta_0 - \delta_1 X - \delta_2 W).$$

Efficiency bound for $\hat{\theta}$ in a CMR model

The semi-parametric efficiency bound for estimating θ in a conditional-moment-restriction model (Chamberlain, 1987)

$$\mathbb{E}[h(Y, Z, X, \theta) \mid X] = 0,$$

is

$$\text{l.b.}(\theta) := (\mathbb{E} J'(X) V^{-1}(X) J(X))^{-1},$$

where

$$J(X) := \partial_\theta \mathbb{E}[h(Y, Z, X, \theta) \mid X],$$
$$V(X) := \mathbb{E}[h(Y, Z, X, \theta) h'(Y, Z, X, \theta) \mid X].$$

Understanding ρ

$$\begin{aligned}&\rho(Y,Z,X,\theta^*)\\&\quad= g(Y,Z,X,\theta^*) + \left(\frac{D}{\pi} - 1\right)\left[g(Y,Z,X,\theta^*) - \mu(Z,X,\theta^*)\right]\\&\qquad := g(Y,Z,X,\theta^*) + \varphi(Y,Z,X,\theta^*),\end{aligned}$$

where $\mathbb{E}[\varphi(Y,Z,X,\theta) \mid X] = 0$ for all θ and $\mathbb{E}[g(Y,Z,X,\theta^*)\varphi(Y,Z,X,\theta^*)] = 0$.

φ is an uninformative penalty to g for not observing all data. Among all functions $\psi(X,Z,\theta^*)$, μ minimises the penalty variance $\mathbb{E}\left[\left(\frac{D}{\pi} - 1\right)(g - \psi) \mid X\right]^2$.

$$\mathbb{E}[\partial_\theta \rho(Y,Z,X,\theta^*) \mid X] \overset{\text{MAR}}{=} \mathbb{E}[\partial_\theta g(Y^*,Z,X,\theta^*) \mid X]$$

$$\mathbb{E}[\partial_{\pi,\mu} \rho(Y,Z,X,\theta^*) \mid X] = 0$$

Bounds for different moment conditions

Function	**Bound for estimating** θ^*
$g(Y^*,X,Z,\theta^*)$	$\left(\mathbb{E}(\partial_\theta g \mid X)'(\mathbb{E}[gg' \mid X])^{-1}\mathbb{E}(\partial_\theta g \mid X)\right)^{-1}$ $= \left(J'(\mathbb{E}[gg' \mid X])^{-1}J\right)^{-1},$
$\frac{D}{\pi(X,Z)}g(Y,X,Z,\theta^*)$	$\left(\mathscr{J}'_{\varpi_*}(\mathbb{E}[\frac{Dgg'}{\pi^2} \mid X])^{-1}\mathscr{J}_{\varpi_*}\right)^{-1}$ $\geq \left(J'(\mathbb{E}[\frac{Dgg'}{\pi^2} \mid X])^{-1}J\right)^{-1},$ $\mathscr{J}_{\varpi_*}$ is a 'noisy' version of J due to π
$\rho(Y,X,Z,\theta^*)$	$\left(J'(\mathbb{E}[\rho\rho' \mid X])^{-1}J\right)^{-1}$ $= \left(J'(\mathbb{E}[\frac{Dgg'}{\pi^2} \mid X] - \mathbb{E}[\frac{1-\pi}{\pi}\mu\mu'])^{-1}J\right)^{-1}$ $\leq \left(J'(\mathbb{E}[\frac{Dgg'}{\pi^2} \mid X])^{-1}J\right)^{-1}$

In the bound associated with ρ, J appears instead of $\mathscr{J}_{\varpi_*}$ because $\mathbb{E}[\partial_{\pi,\mu}\rho(Y,Z,X,\theta^*) \mid X] = 0$.

Empirical likelihood (EL)

Empirical likelihood solves the problem:

$$\max_{p_1,\ldots,p_n,\lambda,\nu} \sum_{i=1}^{n} \log p_i - \lambda \sum_{i=1}^{n} (p_i X_i - \mu) - \nu \sum_{i=1}^{n} (p_i - 1),$$

where μ is the mean *imposed* on the data.

Solution:

$$\underset{\text{(numerically)}}{\hat{\lambda} \text{ solves}} \sum_{i=1}^{n} \frac{X_i - \mu}{1 + \lambda(X_i - \mu)} = 0, \qquad \hat{p}_i = \frac{1}{n} \frac{1}{1 + \hat{\lambda}(X_i - \mu)}.$$

$\lambda \neq 0$ gives the 'distortion' of the probabilities $1/n$ to satisfy the moment condition $\sum_{i=1}^{n} p_i X_i = \mu$.

EL as objective function

Assume $\exists \mu_0 \colon \mathbb{E}[h(X, \mu_0)] = 0$, e. g. $h(X, \mu) = X - \mu$,

$$\hat{\mu} = \arg\max_{\mu} \left[\max_{\substack{p_1, \ldots, p_n, \\ \lambda, \nu}} \sum_{i=1}^{n} \log p_i - \lambda \sum_{i=1}^{n} h(X_i, \mu) - \nu \sum_{i=1}^{n} p_i - 1 \right],$$

$$= \arg\max_{\mu} \left[\max_{\lambda} - \sum_{i=1}^{n} \log\bigl(1 + \lambda h(X_i, \mu)\bigr) \right]$$

$\mathscr{R}(\mu_0) = \prod_{i=1}^{n} n\hat{p}_i$

μ	$\log(\mathscr{R})$	λ
0.12	−0.3297	0.267
0.24	−0.0816	0.142
0.36	**−0.0004**	**−0.011**
0.48	−0.1167	−0.185
0.60	−0.4442	−0.359

Empirical likelihood summary

- EL is a nonparametric method for testing and estimating (Owen, 1988)
- EL 'imposes' a moment restriction on the parameters and data
- EL based on unconditional moment restrictions are equivalent to optimally weighted GMM
- Parametric restrictions can be tested using a nonparametric version of Wilk's theorem (Qin & Lawless, 1994):

$$\mathrm{ELR} := -2 \log(\mathscr{R}(\mu_0)) \xrightarrow{d} \chi^2_{\dim \mu_0}$$

- ELR statistics do not need to be studentised

Smooth empirical likelihood (SEL)

(A, X): generic vector of data, $\{(A_i, X_i)\}_{i=1}^n$: random sample. Kitamura, Tripathi & Ahn (2004): for a model defined by a *conditional* moment restriction $\mathbb{E}[h(A, \theta) \mid X] = 0$, carry out EL for each conditioning X_i, smooth to obtain $\hat{p}_{ij}$:

$$\mathbb{P}(A_j \mid X_i) := \hat{p}_{ij} = w_{ij} \frac{1}{1 + \hat{\lambda}_i' h(A_j, \theta)}$$

with $w_{ij} = \frac{K_b(X_i - X_j)}{\sum_{k=1}^n K_b(X_i - X_k)}$, where $K(\cdot)$ is a 2nd-order kernel and $\hat{\theta}_{\mathrm{SEL}}$ solves the problem

$$\max_\theta \sum_{i=1}^n \sum_{j=1}^n w_{ij} \log \hat{p}_{ij} = \max_\theta \left[-\sum_{i=1}^n \max_{\lambda_i} \sum_{j=1}^n w_{ij} \log\bigl(1 + \lambda_i' h(A_j, \theta)\bigr) \right]$$

Empirical-likelihood-ratio conf. intervals

Let $\mathscr{R}(\mu_0) := \prod_{i=1}^{n} n\hat{p}$. Then,

$$\mathrm{ELR} := -2 \log \mathscr{R}(\mu_0) \xrightarrow{d} \chi^2_{(1)}.$$

ELR test: reject $\mathscr{H}_0 : \mu = \mu_0$ at the α level if $\mathrm{ELR} > Q_{\chi^2_{\dim \mu_0}}(1 - \alpha)$.

In the example above, $\mu_0 = 0$, $\mathrm{ELR}(0) = 1.43$, $Q_{\chi^2_1}(0.95) = 3.84$ – do not reject $\mathscr{H}_0$.

EL and estimating equations

Qin and Lawless (1994): given a model defined by estimating equations expressed as *unconditional* moment conditions $\mathbb{E}[h(A, \theta)] = 0$, where A: generic vector of data, $\{A_i\}_{i=1}^n$: random sample, an estimator $\hat{\theta}$ is given by

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta}\left[\max_{\lambda} \sum_{i=1}^{n} \log p_i\right] \\ &= \arg\max_{\theta}\left[\max_{\lambda} - \sum_{i=1}^{n} \log(1 + \lambda' h(A_i, \theta))\right]\end{aligned}$$

Under suitable regularity conditions,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V), \quad V = [\mathbb{E}(\partial_\theta h)'(\mathbb{E}hh')^{-1}\mathbb{E}(\partial_\theta h)]^{-1}$$

Efficiency of SEL

Why is the SEL estimator semiparametrically efficient?

$$-\frac{1}{n}\nabla_\theta^2 \mathrm{SEL}(\tilde{\theta})\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} B_i + o_p(1)$$

where

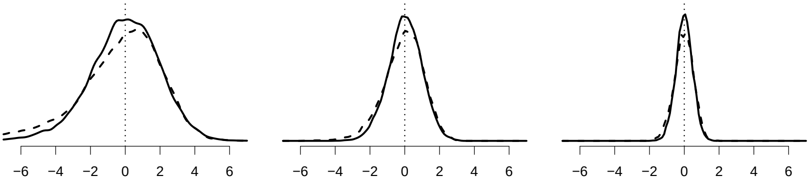
$$B_i := \left(\sum_{j=1}^{n} w_{ij}\partial_\theta g(A_j, \theta_0)\right)' \hat{V}(A_i, \theta_0)^{-1}\left(\sum_{j=1}^{n} w_{ij} g(A_j, \theta_0)\right)$$

$$\hat{V}(A_i, \theta_0) = \sum_{j=1}^{n} w_{ij} g(A_j, \theta_0) g(A_j, \theta_0)'$$

$$-\frac{1}{n}\nabla_\theta^2 \mathrm{SEL}(\tilde{\theta}) \xrightarrow{\mathbb{P}} (\mathbb{E}J'(X)V^{-1}(X)J(X))$$

Estimator distribution

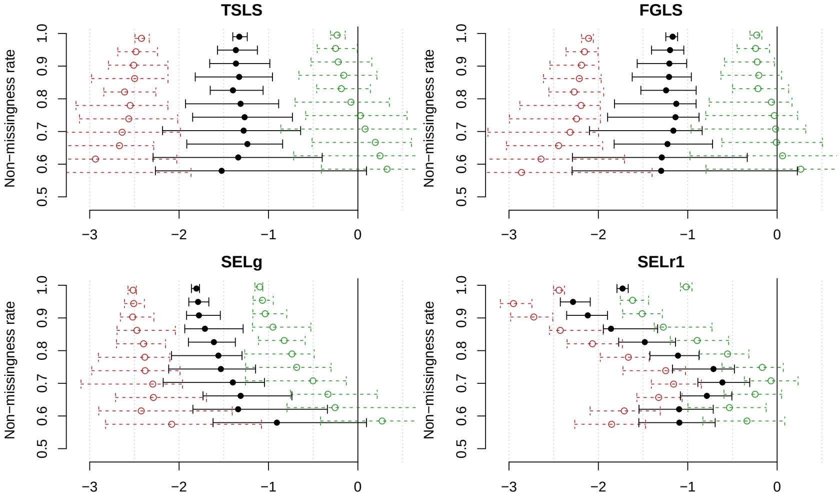
The smoothed density of the centred slope estimator (solid for VS, dashed for FS efficient) is shown below.



n = 500 n = 2000 n = 8000

The full-sample estimator is more tightly concentrated around the true value, has thinner tails, and looks Gaussian.

Application: conf. intervals (MOREKIDS)


TSLS
FGLS
SELg
SELr1
Non−missingness rate