Missing Endogenous Variables
in Conditional-Moment-Restriction Models

Antonio Cosma **Andreï V. Kostyrka** Gautam Tripathi

DSEF Jamboree
September 23, 2020
Universitéit Lëtzebuerg

UNIVERSITÉ DU LUXEMBOURG
Department of Economics
and Management (DEM)

Presentation structure

1. Showing common examples of problems with missing data in endogenous variables.
2. Discussing parameter identification.
3. Deriving the efficiency bounds.
4. Specifying the efficient estimator.
5. Presenting the simulation results.

Motivation behind our work

- Standard statistical methods have been developed to analyse rectangular data sets.
- In practice, often data entries for dependent and endogenous variables are missing:
  - Households might refuse to report income;
  - Individuals might report nothing instead of nutrient intake in human productivity studies.
- Not so many works have dealt with this issue recently.
  - Chen, Hong, Tarozzi (2008, *AoS*), Graham (2011, *Ecta*) consider only unconditional restrictions.
  - Hristache, Patilea (2017, *Biometrika*) provide theoretical results (but no efficiency bound) for CMR models.
- **Our contribution:** derive the semi-parametric efficiency bound in CMR models with missing endogenous variables and propose an estimator that attains it.

Basic one-sample model

- Many econometric models can be written in the form

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

  where θ^* is the parameter vector of interest.
  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.
- However, the observed version of Y^* is

$$Y \stackrel{\text{def}}{=} DY^* + (1 - D)\mathfrak{m},$$

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

Estimation objective

- Using observations $(Y_i, D_i, X_i, Z_i)_{i=1}^n$, efficiently estimate θ^* and average partial effects w. r. t. X.
- An example data set can be seen on the right.

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	$\mathfrak{m}$	2.09
−0.98	0.38	1	−0.52	−0.52
0.91	0.38	0	$\mathfrak{m}$	1.37
1.92	0.39	0	$\mathfrak{m}$	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	$\mathfrak{m}$	2.53
−0.48	0.66	1	−0.28	−0.28
1.18	0.69	0	$\mathfrak{m}$	1.70

Examples of the one-sample problem

- IV regression with missing outcomes:

  $$\mathbb{E}(\mathit{wage}^* - \alpha^* - X'_{\text{incl}}\beta^* - \gamma^* \mathit{education} \mid X_{\text{incl}}, X_{\text{excl}}) = 0$$

  Example: estimating returns to education.
- IV regression with missing endogenous explanatory variables:

  $$\mathbb{E}(\mathit{ART} - \alpha^* - X'_{\text{incl}}\beta^* - \gamma^* \mathit{mental\ health}^* \mid X_{\text{incl}}, X_{\text{excl}}) = 0$$

  Example: estimating the effect of mental health status (not always reported) on anti-retroviral treatment adherence.
  Example: estimating the effect of infant health (often unknown) on labour market outcomes.

Identifying assumptions

Recall the model in the one-sample case:

$$\mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X] = 0, \quad D = 1 \text{ if } Y = Y^*, \quad D = 0 \text{ if } Y = \mathfrak{m}$$

We assume **missingness at random** (MAR):

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

The missingness indicator D is conditionally independent of Y^* and the always observed endogenous variables Z.

Then, define $\pi(X) \stackrel{\text{def}}{=} \mathbb{E}(D \mid X)$ to be the **unknown propensity score function** with $0 < \pi(X) < 1$.

Identification

Whilst $g(Y^*, Z, X, \theta^*)$ cannot be evaluated since Y^* is not observed, it can be shown that, by the MAR assumption,

$$\begin{aligned} 0 = \mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X] &= \mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X, D = 1] \\ &= \mathbb{E}\left[\frac{Dg(Y^*, Z, X, \theta^*)}{\pi(X)} \,\middle|\, X\right] \end{aligned}$$

Under MAR, the **validation sample** (VS, where $D = 1$) alone is enough to identify θ^*.

Problem? Efficiency is lost.

Main theoretical result

The efficient estimator must use the information from the full sample. The model yielding such an estimator is based on the **transformed moment function** ρ:

$$\mathbb{E}[\rho(\mathcal{A}, \theta^*, \pi, \mu) \mid X] = 0, \quad \text{where}$$

$$\rho(\mathcal{A}, \theta^*, \pi, \mu) \stackrel{\text{def}}{=} \frac{Dg(Y, Z, X, \theta^*)}{\pi(X)} - \mu(Z, X, \theta^*)\Big(\frac{D}{\pi(X)} - 1\Big),$$

$$\mathcal{A} \stackrel{\text{def}}{=} (Y, Z, D, X), \quad \mu(Z, X, \theta^*) \stackrel{\text{def}}{=} \mathbb{E}[g(Y^*, Z, X, \theta^*) \mid Z, X].$$

$$\text{l.b.}(\theta^*)\big|_\rho \leq \text{l.b.}(\theta^*)\big|_g$$

An asymptotically efficient estimator **beats** (in the sense of Chamberlain, 1987, *JoE*) **any estimator** constructed using the validation sample alone!

Implications for estimation

- The efficiency gains measured by the ratio $\text{l.b.}(\theta^*)|_g$ / $\text{l.b.}(\theta^*)|_\rho$ are due to the presence of the non-missing endogenous variables.
  - If there are no endogenous variables in the model that are not missing, i. e. all of the endogenous variables in the model are missing, then estimating θ^* using the VS alone is asymptotically efficient.
- $\pi(X)$ can be fully unknown, and should always be estimated non-parametrically even if it is known up to a finite-dimensional parameter or fully known.
  - Estimating π and μ non-parametrically does not affect the asymptotic distribution of the estimator.

Estimation method

- Smoothed Empirical Likelihood (SEL, proposed by Kitamura, Tripathi & Ahn, 2004, *Ecta*) extends the Empirical Likelihood, a non-parametric method for testing and estimating (Owen, 1988, *Biometrika*).
- Parametric restrictions can be tested using a non-parametric version of Wilks' theorem (Qin and Lawless, 1994, *Ann. Stat.*). EL ratio statistics do not need to be explicitly studentised.
- SEL extends the properties of EL to estimating model characterised by conditional moment restrictions (Kitamura & Tripathi, 2003, *Ann. Stat.*), and SEL-based estimators attain the semi-parametric efficiency bounds.

Implementation of our estimator

In order to take into account conditioning, we construct kernel weights (with bandwidth b)

$$w_{ij} \stackrel{\text{def}}{=} \frac{K_b(X_i - X_j)}{\sum_{k=1}^{n} K_b(X_i - X_k)}, \quad i, j = 1, \ldots, n.$$

The SEL estimator solves the optimisation problem:

$$\max_{\theta} - \sum_{i=1}^{n} \max_{\lambda_i} \sum_{j=1}^{n} w_{ij} \log\big(1 + \lambda_i' \hat{\rho}(\mathcal{A}_j, \theta)\big).$$

Numerical optimisation can be used to solve both maximisation problems.

Non-parametric imputation

Since $\rho(Y, Z, D, X, \theta, \pi, \mu)$ depends on unknown functions, $\pi(X) \stackrel{\text{def}}{=} \mathbb{E}(D \mid X)$ and $\mu(X, Z, \theta) \stackrel{\text{def}}{=} \mathbb{E}[g(Y^*, Z, X, \theta) \mid Z, X]$, they can be estimated via kernel regression methods (i. e. Nadaraya—Watson estimator):

$$\hat{\pi}(X) \stackrel{\text{def}}{=} \frac{\sum_{k=1}^{n} D_k K_{c_1}(X_k - X)}{\sum_{k=1}^{n} K_{c_1}(X_k - X)},$$

$$\hat{\mu}(Z, X, \theta) \stackrel{\text{def}}{=} \frac{1}{\hat{\pi}(X)} \frac{\sum_{k=1}^{n} D_k g(Y_k, Z_k, X_k, \theta) K_{c_2}(Z_k - Z, X_k - X)}{\sum_{k=1}^{n} K_{c_2}(Z_k - Z, X_k - X)},$$

where $K_{c_1}(\cdot), K_{c_2}(\cdot)$ are kernel functions and c_1, c_2 are bandwidths.

Inference

- The SEL approach provides a convenient unified environment for testing hypotheses about θ^* using the likelihood ratio: $\mathrm{LR}(\theta) \stackrel{\text{def}}{=} 2[\mathrm{SEL}(\hat{\theta}) - \mathrm{SEL}(\theta)]$.
- Consider any parametric restriction $\mathcal{H}_0 \colon R(\theta^*) = 0$. Then, maximise the SEL under the constraint: $\hat{\theta}_R \stackrel{\text{def}}{=} \underset{\theta \colon R(\theta)=0}{\operatorname{argmax}} \mathrm{SEL}(\theta)$.
- Reject $\mathcal{H}_0$ if $\mathrm{LR}(\hat{\theta}_R) > Q_{\chi^2_{\dim R}}(\alpha)$ for the desired level α.
- The LR statistic can be inverted to obtain asymptotically valid confidence regions: $\{\theta \colon \mathrm{LR}(\theta) \leq Q_{\chi^2_{\dim \theta}}(\alpha)\}$.

Simulation results (discrete design)

$$Y^* = 1 + 1 \cdot Z + U\sigma(X), \quad \mathbb{E}(U \mid X) = 0$$

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

$$X \sim \text{Bernoul}(0.6), \; Z = \mathbb{I}(X + V > 0)$$

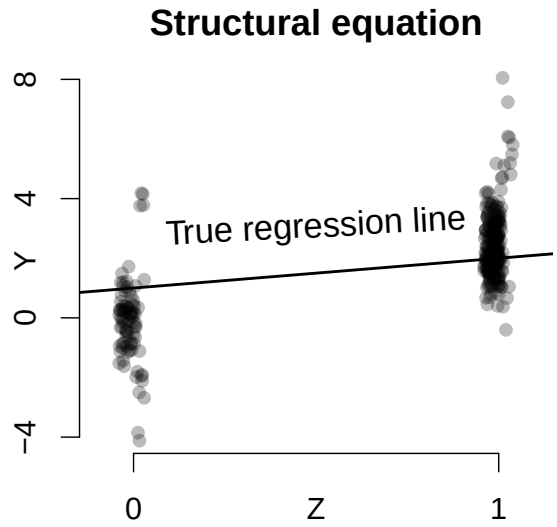
$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right]$$

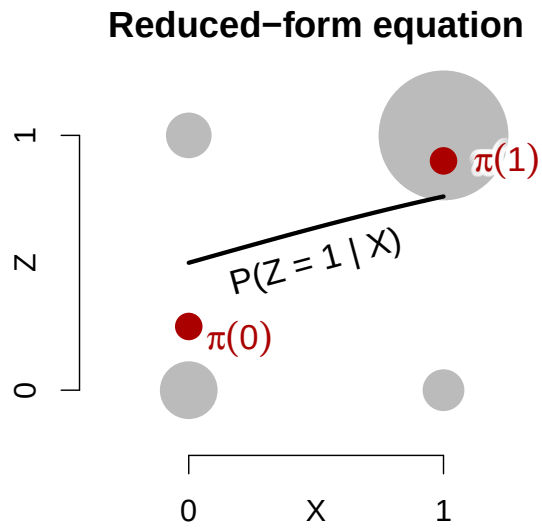
$$\pi(X) = 0.9X + 0.25(1 - X)$$

$$\sigma^2(X) = X + 16(1 - X)$$

Max. gains: 31%, missing: 36%

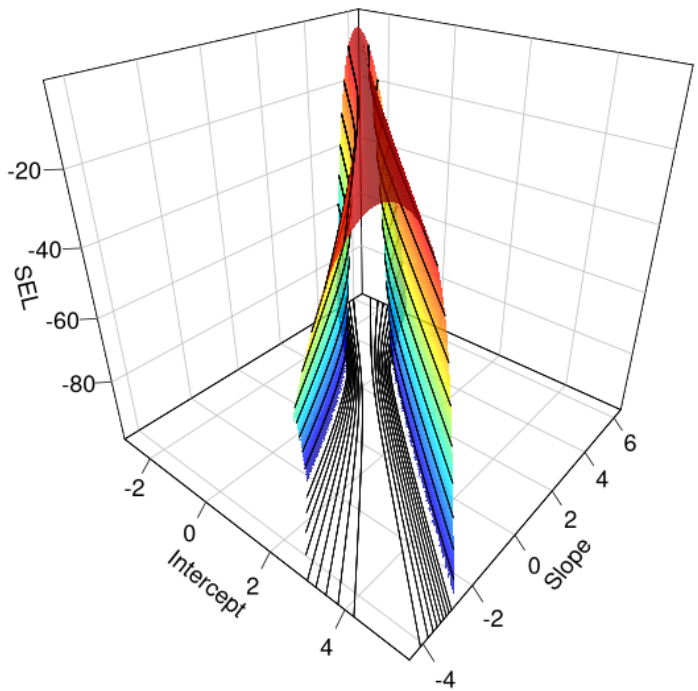
Example: effect of treatment (Z) with eligibility indicator X.

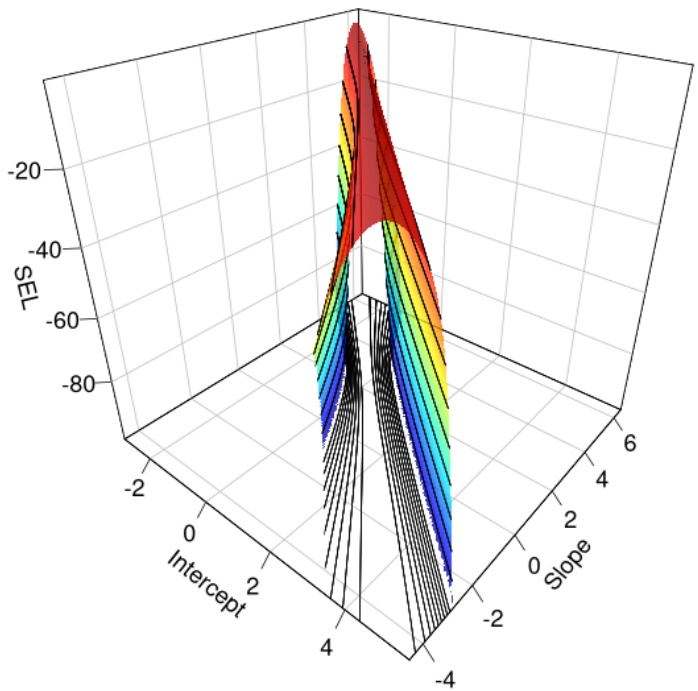




Objective function behaviour

SEL(θ) might look similar in the VS and full sample...

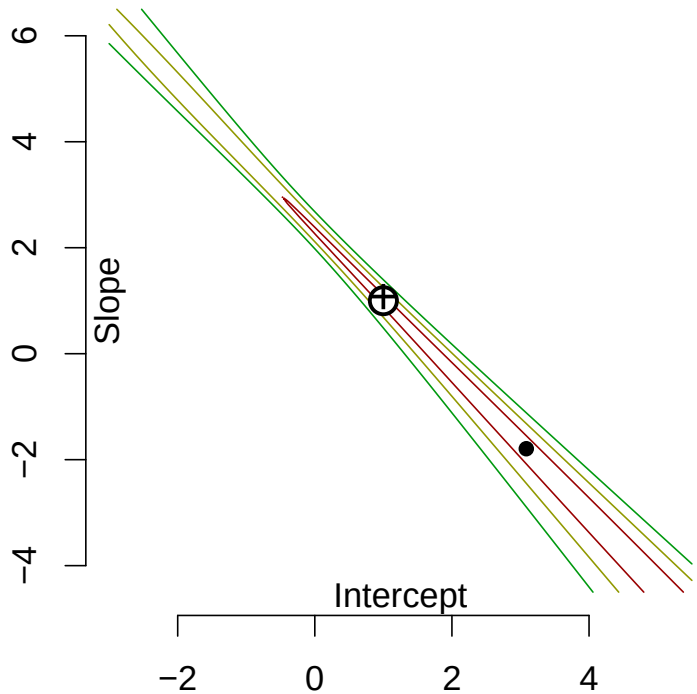




VS only $n = 500$ Full sample

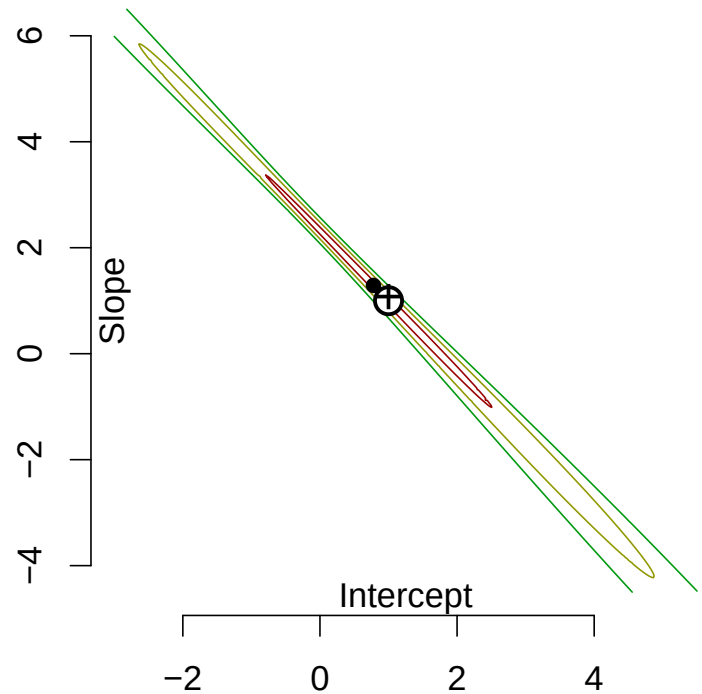
SEL-based confidence regions

...but the implications for inference are huge!
$\oplus$ denotes the true value, $\bullet$ denotes the estimate.
Levels: 50%, 95%, 99.9%.



VS only

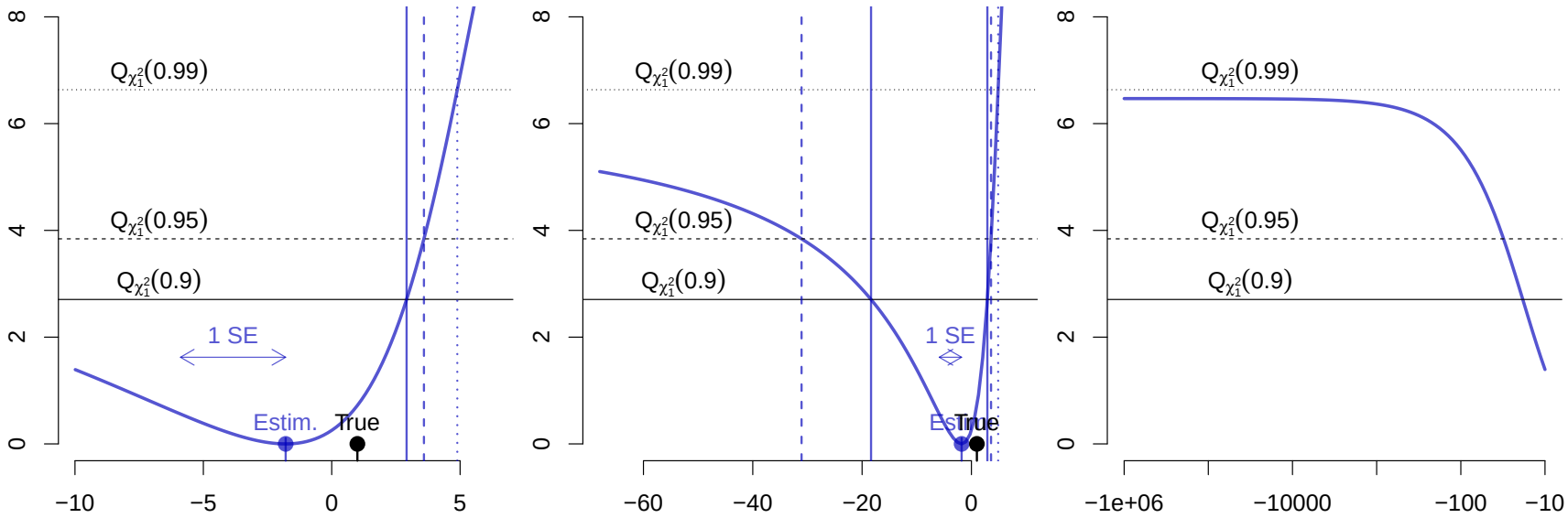
$n = 500$



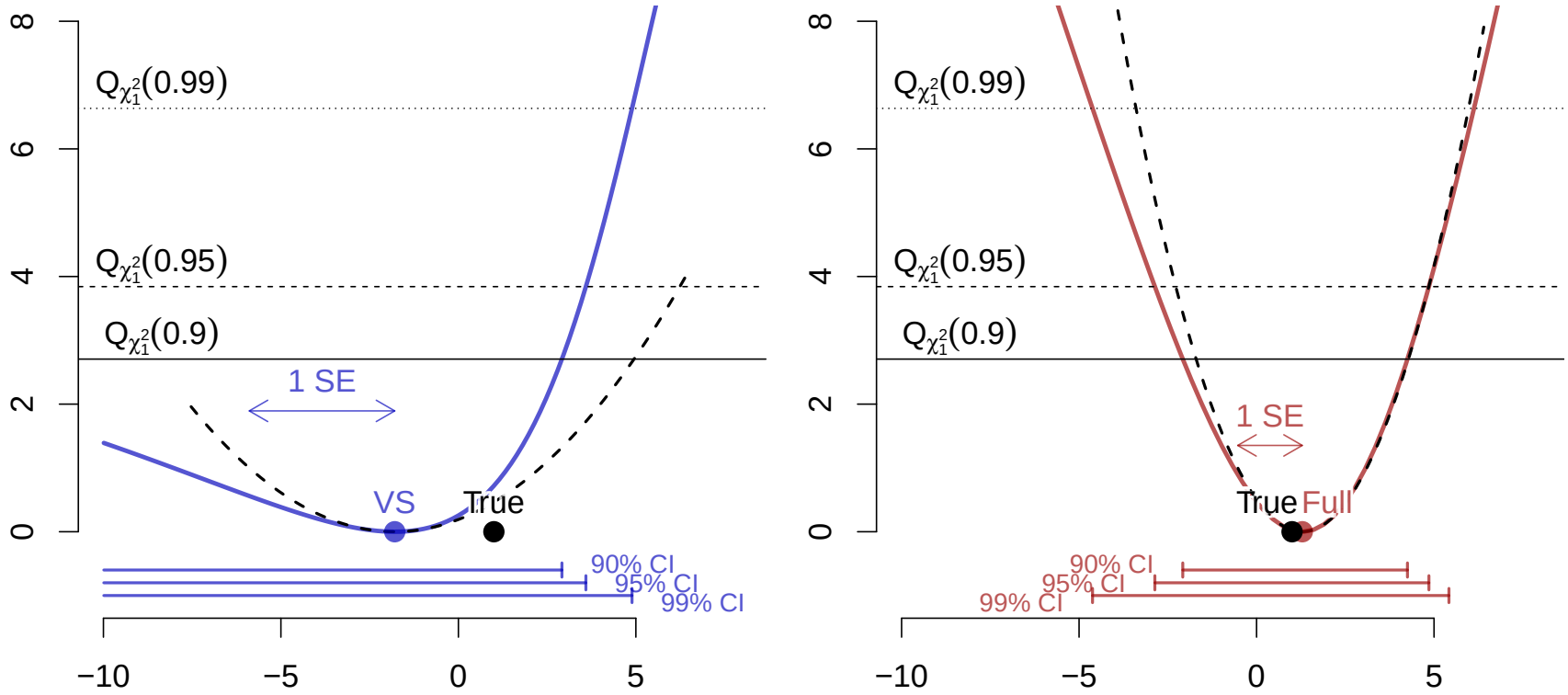
Full sample

LR statistic for the slope (VS only)

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



Profile LR comparison



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

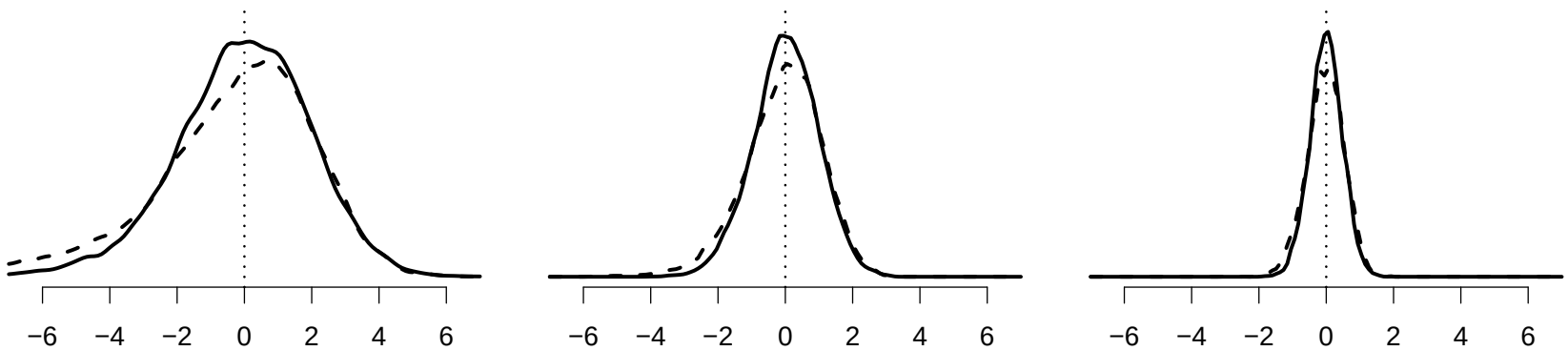
Monte-Carlo simulation summary

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

n	Est.	Bias		SD	Med. AD ratio	Gains
		Med.	Mean			
500	FS	0.049	−0.047	2.018	1.000	546%
	VS	−0.032	−0.570	5.100	1.163	
2000	FS	0.032	0.013	0.969	1.000	44.6%
	VS	0.013	−0.078	1.162	1.170	
8000	FS	0.001	−0.002	0.479	1.000	35.2%
	VS	−0.003	−0.027	0.556	1.171	

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.

Confidence interval summary

		Coverage probability			Bounded
n	Est.	Nominal	Empirical	Length	intervals (%)
500	FS	.90	.905	7.1	100
		.95	.953	8.7	100
		.99	.991	12.6	100
	VS	.90	.898	16.0	96.9
		.95	.953	24.6	94.2
		.99	.991	82.5	83.9
2000	FS	.90	.896	3.2	100
		.95	.951	3.9	100
		.99	.990	5.2	100
	VS	.90	.899	3.9	100
		.95	.950	4.8	100
		.99	.990	6.7	100

Further work

- Run simulations with continuously distributed regressors.
- Handle the bandwidth issue for non-parametric prediction via smoothing.
- Show consistency, efficiency, and asymptotic normality of the estimator.
- Apply the method to a real data set.

Conclusions

- We show that if at least one of endogenous variables in the models contains missing values, and if not all endogenous variables are missing, then there are efficiency gains compared to the classical complete-case approach, and derive the bound.
- We propose an estimator that attains said efficiency bound.
- We test its performance in practice and find that it yields empirical gains close to theoretically expected ones.

Thank you for your attention!

Questions?