A Robust Instrumental Variable Estimation Method with Its Applications

Yiqi LIN

Department of Statistics, The Chinese University of Hong Kong
with F. Windmeijer, X. Song, and Q. Fan

2022-2023 Faculty Postgraduate Research Day, 8 Feb 2023

Endogeneity issue

A linear model:

$$Y = \beta D + \epsilon. \tag{1}$$

Y is the outcome variable, D is the treatment variable, β is the treatment effect. Under the gold standard of randomized controlled trials, $E(\epsilon|D) = 0$ (exogeneity). But exogenous treatment is often violated in observational studies due to:

- Omitted variables (unmeasured confounders): $\epsilon = \boldsymbol{X}_2 + u$, $E(u'D) = 0$, and $cov(D, \boldsymbol{X}_2) \neq \mathbf{0}$
- measurement error in D: $D^{ob} = D + u, E(u) = 0$.
- Sample selection, simultaneous equations, reverse causality, private information...

In above situations, $E(\epsilon|D) \neq 0$. The treatment variable D is endogenous and the simple OLS cannot provide the consistent estimate:

$$\hat{\beta}_{OLS} \xrightarrow{p} \beta \tag{2}$$

- **One common solution is to use instrumental variable (IV).**

Requirement of IVs

A good IV should satisfy the following conditions, also illustrated as follows.

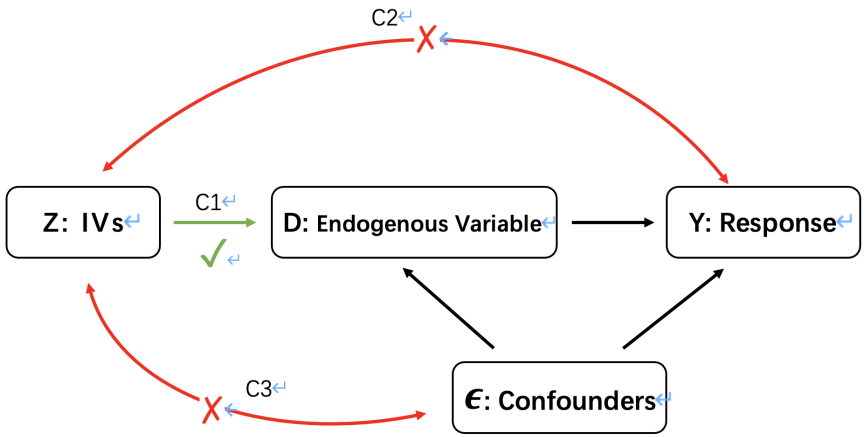

Figure: Illustration of Validity and Relevance.

IVs Requirements

- C1 Relevance Condition: related to exposure (may be strong or weak).
- C2 Exogenous Condition: not related to unmeasured confounders that affect the exposure and the outcome.
- C3 Exclusion Restriction: have no direct pathway to the outcome.

Practical and theoretical issues with IV

While in theory we know what qualifies as a good IV, in practice it is not very easy to find good IVs.

- Weak and invalid IVs are common (Stock et al. 2002, Andrews et al., 2018).
- Usually no economic theory as guidance on the identity of IVs.
- Research in gene data (mendelian randomized study), SNPs usually serves as IVs, But only few knowledge about whether some SNPs are valid or not is available.
- Model identification is the first order problem when the validity is unknown.

Generally, we have a set of candidates of IVs, but we don't know which is valid and useful.

Research plan and our contribution

What we do:

- Propose an estimator robust to Weak and Invalid instruments for Treatment effects (WIT). The model with unknown invalid IV and all weak IV is arguably the most difficult case in theory yet common in practice. The "sparsest rule" assumption has merits in theory and practice regarding the research design.
- Provide algorithm for the proposed estimator based on non-convex penalty functions. Moreover, our computation algorithm is in alignment with the identification condition.
- Establish theorems for the identification problem and asymptotic behavior of our estimator.

Contributions:

- Address the fundamental issue of linear IV model identification with unknown invalid IVs.
- Show the impossibility result of the necessary condition.
- Provide solutions to unify model identification and model selection consistency.
- Utilize weak instruments in selecting valid IVs.

Model

A valid instrument $\boldsymbol{Z}_j$ should not have a direct effect on the response and unmeasured confounders

Model

Assuming the linear functional form between treatment effects $\boldsymbol{D}_i$ and instruments $\boldsymbol{Z}_{i.}$,

$$\begin{aligned} Y_i &= D_i\beta^* + \boldsymbol{Z}_{i.}^\top \boldsymbol{\alpha}^* + \epsilon_i \\ D_i &= \boldsymbol{Z}_{i.}^\top \boldsymbol{\gamma}^* + \eta_i. \end{aligned} \tag{3}$$

Definition

- Relevant IV (satisfies C1): if $\gamma_j^* \neq 0, j = 1, 2, \ldots, p$.
- Valid IV (satisfies C2 and C3): if $\alpha_j^* = 0, j = 1, 2, \ldots, p$.
- Define the valid IV set $\mathcal{V}^* = \{j : \alpha_j^* = 0\}$ and invalid IV set $\mathcal{V}^{c*} = \{j : \alpha_j^* \neq 0\}$.

Identifiability of Model

Mixture of valid and Invalid IVs

We must have some valid IVs, but usually we don't know exactly which ones are. That means we are facing a mixed set of IVs, some of which are valid and some are not.

Example (with $p = 3$)

$$\text{(Structural equation) } \boldsymbol{Y} = \boldsymbol{D}\beta^* + \boldsymbol{Z}_1\boldsymbol{\alpha}_1^* + \boldsymbol{Z}_2\boldsymbol{\alpha}_2^* + \boldsymbol{\epsilon} \Rightarrow \boldsymbol{\alpha}^* = (\boldsymbol{\alpha}_1^*, \boldsymbol{\alpha}_2^*, 0)$$

$$\text{(First stage equation) } \boldsymbol{D} = \boldsymbol{Z}_1\boldsymbol{\gamma}_1^* + \boldsymbol{Z}_2\boldsymbol{\gamma}_2^* + \boldsymbol{Z}_3\boldsymbol{\gamma}_3^* + \boldsymbol{\eta}$$

Then we rearrange first stage equation:

$$\boldsymbol{Z}_1\boldsymbol{\gamma}_1^* = \boldsymbol{D} - \boldsymbol{Z}_2\boldsymbol{\gamma}_2^* - \boldsymbol{Z}_3\boldsymbol{\gamma}_3^* - \boldsymbol{\eta}$$

$$\Rightarrow \boldsymbol{Z}_1\boldsymbol{\alpha}_1^* = \boldsymbol{Z}_1\boldsymbol{\gamma}_1^*\left(\frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*}\right) = \boldsymbol{D}\frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*} - \boldsymbol{Z}_2\boldsymbol{\gamma}_2\frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*} - \boldsymbol{Z}_3\boldsymbol{\gamma}_3^*\frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*} - \boldsymbol{\eta}\frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*}$$

$$\Rightarrow \boldsymbol{Y} = \boldsymbol{D}\beta^* + \left(\boldsymbol{D}\frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*} - \boldsymbol{Z}_2\boldsymbol{\gamma}_2\frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*} - \boldsymbol{Z}_3\boldsymbol{\gamma}_3^*\frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*} - \boldsymbol{\eta}\frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*}\right) + \boldsymbol{Z}_2\boldsymbol{\alpha}_2^* + \boldsymbol{\epsilon}$$

$$\Rightarrow \boldsymbol{Y} = \boldsymbol{D}\left(\beta^* + \frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*}\right) + \boldsymbol{Z}_2\left(\boldsymbol{\alpha}_2^* - \frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*}\right) + \boldsymbol{Z}_3\left(-\frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*}\right) + \left(\boldsymbol{\epsilon} - \frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*}\boldsymbol{\eta}\right)$$

Then it forms a new DGP: $\tilde{\beta} = \beta^* + \frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*}$, $\tilde{\boldsymbol{\alpha}} = (0, \boldsymbol{\alpha}_2^* - \frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*}, -\frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*})$ and $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} - \frac{\boldsymbol{\alpha}_1^*}{\boldsymbol{\gamma}_1^*}\boldsymbol{\eta}$.

Identifiability of Model (Cont')

Let $\mathcal{Q}$ to be the collections of all possible population-level solution of $\boldsymbol{\alpha}$ under the constraint, some components need to be 0.

The Sparsest Rule

$$\boldsymbol{\alpha}^* = \text{argmin}_{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}} \|\boldsymbol{\alpha}\|_0$$

Consider the general penalized TSLS estimator with penalty on $\boldsymbol{\alpha}$:

$$\left(\hat{\boldsymbol{\alpha}}^{\text{pen}}, \hat{\beta}^{\text{pen}}\right) = \underset{\boldsymbol{\alpha},\beta}{\text{argmin}} \underbrace{\frac{1}{2n}\|P_Z(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\alpha} - \boldsymbol{D}\beta)\|_2^2}_{(I)} + \underbrace{p_\lambda^{\text{pen}}(\boldsymbol{\alpha})}_{(II)}. \tag{4}$$

They have the two different functions: (I) approximate the structure of $\mathcal{Q}$ and (II) imposed to ensure sparsity structure in $\boldsymbol{\alpha}$.

Dual Form View (Aligned with Sparsest Rule)

$$\left(\hat{\boldsymbol{\alpha}}^{\text{opt}}, \hat{\beta}^{\text{opt}}\right) = \underset{\boldsymbol{\alpha},\beta}{\text{argmin}} \|\boldsymbol{\alpha}\|_0 \quad \text{s.t.} \quad \underbrace{\|P_Z(\boldsymbol{Y} - \boldsymbol{D}\beta - \boldsymbol{Z}\boldsymbol{\alpha})\|_2^2 < \delta}_{\text{only possible in } \mathcal{Q}}.$$

WIT Estimator

Proposition 1 (The proper Surrogate Sparsest penalty)

If $p_\lambda^{\text{pen}}(\boldsymbol{\alpha})$ is surrogate sparsest rule in the sense of that it gives sparse solutions and

$$\boldsymbol{\alpha}^* = \underset{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}}{\operatorname{argmin}} \|\boldsymbol{\alpha}\|_0 = \underset{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}}{\operatorname{argmin}} p_\lambda^{pen}(\boldsymbol{\alpha}),$$

then $p_\lambda^{pen}(\cdot)$ must be concave.

We adopt the penalized method framework (10) and deploy a concave penalty in (11), the MCP (Zhang et al. 2010) in particular, which is a nearly unbiased estimator.

WIT Estimator (First Stage: Select Valid IVs)

$$\widehat{\boldsymbol{\alpha}}^{\text{MCP}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2n}\|\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\boldsymbol{\alpha}\|_2^2 + p_\lambda^{\text{MCP}}(\boldsymbol{\alpha}), \quad \hat{\mathcal{V}} = \{j : \widehat{\alpha}_j^{\text{MCP}} = 0\},$$

Selection Consistency

Theorem 3 (Selection Consistency of Valid IVs)

Specify $\kappa(n)$ and $\kappa^c(n)$ in Assumption 5, where $\tilde{\tilde{Q}}_n = Z_{\mathcal{V}^*}^\top (P_Z - P_{Z_{\mathcal{V}^c}}) Z_{\mathcal{V}^*}/n$.

$$\kappa(n) \asymp \underbrace{\sqrt{\frac{\log p_{\mathcal{V}^*}}{n}}}_{T_1} + \underbrace{\frac{p_{\mathcal{V}^*}}{n} \cdot \frac{\|\tilde{\tilde{Q}}_n \gamma^*_{\mathcal{V}^*}\|_\infty}{\gamma^{*\top}_{\mathcal{V}^*} \tilde{\tilde{Q}}_n \gamma^*_{\mathcal{V}^*}}}_{T_2} + \underbrace{|\,\mathrm{Bias}(\hat{\beta}^{\mathrm{TSLS}}_{\mathrm{or}})| \|\bar{\gamma}^*_{\mathcal{V}^{c*}}\|_\infty}_{T_3}, \tag{5}$$

$$\kappa^c(n) \asymp (1+c)\left\{\sqrt{\frac{\log |\mathcal{I}_c|}{n}} + \frac{|\mathcal{I}_c|}{n} \cdot \frac{\|\tilde{\tilde{Q}}^c_n \gamma^*_{\mathcal{I}_c}\|_\infty}{\gamma^{*\top}_{\mathcal{I}_c} \tilde{\tilde{Q}}^c_n \gamma^*_{\mathcal{I}_c}}\right\} + |\,\mathrm{Bias}(\hat{\beta}^{c\ \mathrm{TSLS}}_{\mathrm{or}})| \|\bar{\gamma}^*_{\mathcal{I}^c_c}\|_\infty, \tag{6}$$

Moreover, under assumptions $|\tilde{\alpha}^c_j| > \kappa^c(n)$ and $|\alpha^*_{\mathcal{V}^{c*}}|_{\min} > \kappa(n)$, then

$$\hat{\alpha}^{\mathrm{MCP}} = \underset{\hat{\alpha} \in \mathscr{B}_0(\lambda,\rho)}{\mathrm{argmin}} \|\hat{\alpha}\|_0, \quad \Pr(\hat{\mathcal{V}} = \mathcal{V}^*, \hat{\alpha}^{\mathrm{MCP}} = \hat{\alpha}^{\mathrm{or}}) \xrightarrow{p} 1. \tag{7}$$

1. T_1: Common rate in Lasso/non-convex penalty in high-dim linear regression.
2. T_2 and T_3: Additional difficulty in penalized regression within IV contents:
   - T_2: many IVs risk,
   - T_3: bias in weak IVs problem.

WIT Estimator

WIT Estimator (Second Stage: Estiamte)

$$\hat{\beta}^{\text{WIT}} = \left(\boldsymbol{D}_{\perp}^{\top} \left(\boldsymbol{I} - \hat{\kappa}_{\text{liml}} M_{\boldsymbol{Z}_{\hat{\mathcal{V}}_{\perp}}} \right) \boldsymbol{D}_{\perp} \right)^{-1} \left(\boldsymbol{D}_{\perp}^{\top} \left(\boldsymbol{I} - \hat{\kappa}_{\text{liml}} M_{\boldsymbol{Z}_{\hat{\mathcal{V}}_{\perp}}} \right) \boldsymbol{Y}_{\perp} \right)$$

$$\hat{\kappa}_{\text{liml}} = \lambda_{\min} \left(\{ [\boldsymbol{Y}_{\perp}, \boldsymbol{D}_{\perp}]^{\top} M_{\boldsymbol{Z}_{\hat{\mathcal{V}}_{\perp}}} [\boldsymbol{Y}_{\perp}, \boldsymbol{D}_{\perp}] \}^{-1} \{ [\boldsymbol{Y}_{\perp}, \boldsymbol{D}_{\perp}]^{\top} [\boldsymbol{Y}_{\perp}, \boldsymbol{D}_{\perp}] \} \right)$$

Theorem 4 (Consistency and Asymptotic Normality)

Under same conditions in Theorem 3, with some additional assumptions, we obtain:

1. (Consistency): $\hat{\beta}^{\text{WIT}} \xrightarrow{p} \beta^*$ with $\hat{\kappa}_{\text{liml}} = \frac{1-\upsilon_L}{1-\upsilon_K-\upsilon_L} + o_p(1)$.
2. (Asymptotic normality): $\sqrt{n}(\hat{\beta}^{\text{WIT}} - \beta^*) \xrightarrow{d} \mathcal{N}\left(0, \mu_0^{-2} \left[\sigma_\epsilon^2 \mu_0 + \frac{\upsilon_K(1-\upsilon_L)}{1-\upsilon_K-\upsilon_L} |\boldsymbol{\Sigma}| \right] \right)$.
3. (Consistent variance estimator):

$$\widehat{\text{Var}}(\hat{\beta}^{\text{WIT}}) = \frac{\hat{\boldsymbol{b}}^{\top} \hat{\boldsymbol{\Omega}} \hat{\boldsymbol{b}} (\hat{\mu}_n + L/n)}{-\hat{\mu}_n} \left(\hat{Q}_S \hat{\boldsymbol{\Omega}}_{22} - \boldsymbol{T}_{22} + \frac{\hat{c}}{1-\hat{c}} \frac{\hat{Q}_S}{\hat{\boldsymbol{a}}^{\top} \hat{\boldsymbol{\Omega}}^{-1} \hat{\boldsymbol{a}}} \right)^{-1} \xrightarrow{p} \text{Var}(\sqrt{n} \hat{\beta}^{\text{WIT}}),$$

where $\hat{\boldsymbol{b}} = (1, -\hat{\beta}^{\text{WIT}})$ and $\hat{Q}_S = \frac{\hat{\boldsymbol{b}}^{\top} \boldsymbol{T} \hat{\boldsymbol{b}}}{\hat{\boldsymbol{b}}^{\top} \hat{\boldsymbol{\Omega}} \hat{\boldsymbol{b}}}$.

Simulation (weak IVs)

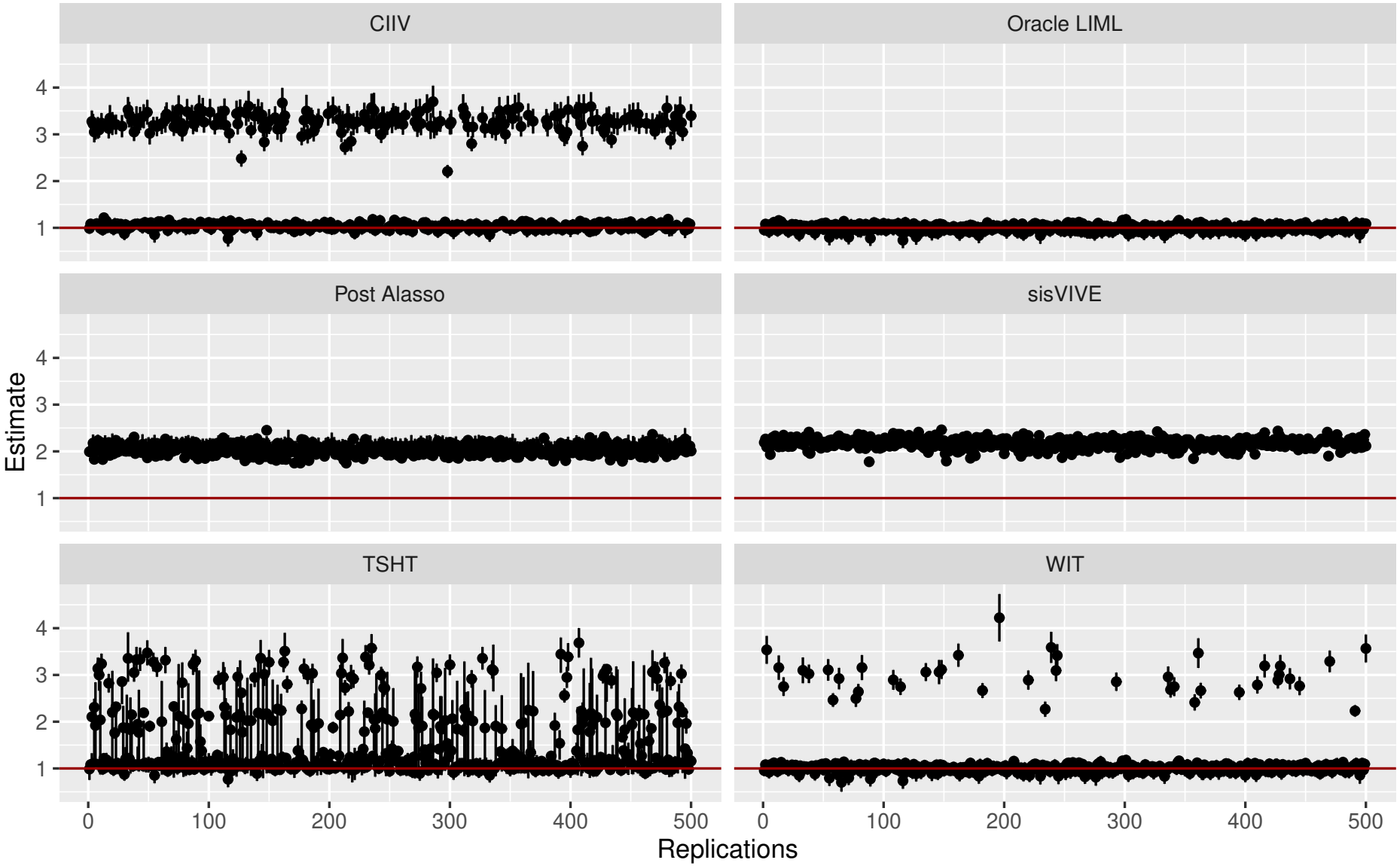

Figure: Scatter plot of estimations of β^* with confidence intervals of Case 1 (IV)

Real Data Application: Trade and growth

- We re-investigate the causal effect of trade on income (FR99) using a more comprehensive and updated data, taking into account that trade is an endogenous variable (it correlates with unobserved common factors driving both trade and growth) and some instruments might be invalid.
- The structural equation considered in FR99 is,

$$\log(Y_i) = \alpha + \beta T_i + \psi S_i + \epsilon_i \tag{8}$$

where for each country i, Y_i is the GDP per worker, T_i is the share of international trade to GDP, S_i is the size of the country, such as area, population, and ϵ_i is the error term.

- FR99 proposed to construct an IV (called a proxy for trade) based on the celebrated gravity theory of trade (Anderson, 1979). The logic of IV validity in aggregate level is that the geographical variables, such as common border and distance between countries, indirectly affect growth through the channel of convenience for trade.

Trade on GDP

Following the same logic, Fan and Zhong (2018) extended the IV set to include more geographic and meteorological variables. The reduced form equation is

$$T_i = \boldsymbol{\gamma}^\top \boldsymbol{Z_i} + \nu_i, \tag{9}$$

where $\boldsymbol{Z_i}$ is a vector of instruments.

Table: Summary statistics of main variables

	Notation	Type	Mean	Std	Median	Min	Max
log(GDP)	$\log(Y)$	Response	10.177	1.0102	10.416	7.463	12.026
Trade	T	Endogenous Variable	0.866	0.520	0.758	0.198	4.128
log(Population)	$\boldsymbol{S}_1$	Control Variable	1.382	1.803	1.480	−3.037	6.674
log(Land Area)	$\boldsymbol{S}_2$	Control Variable	11.726	2.260	12.015	5.680	16.611
$\widehat{T}$ (proxy for trade)	$\boldsymbol{Z}_1$	IV	0.093	0.052	0.079	0.015	0.297
log(Water Area)	$\boldsymbol{Z}_2$	IV	6.756	3.654	7.768	0	13.700
log(Land Boundaries)	$\boldsymbol{Z}_3$	IV	6.507	2.920	7.549	0	10.005
% Forest	$\boldsymbol{Z}_4$	IV	29.89	22.380	30.62	0	98.26
% Arable Land	$\boldsymbol{Z}_5$	IV	40.947	21.549	42.062	0.558	82.560
Languages	$\boldsymbol{Z}_6$	IV	1.873	2.129	1	1	16
Annual Freshwater	$\boldsymbol{Z}_7$	IV	2.190	2.129	2.155	-2.968	8.767

Source: FR99, the World Bank, and CIA world Factbook.

Estimation

Table: Empirical Results of Various Estimators

	$\hat{\beta}\left(\widehat{\text{Var}}^{1/2}(\hat{\beta})\right)$	95% CI	Valid IVs $\widehat{\mathcal{V}}$	Relevant IVs $\hat{\mathcal{S}}$	Sargan Test
OLS	0.413(0.084)	(0.246, 0.581)	-	-	-
FR99	0.673(0.220)	(0.228, 1.117)	-	-	0.999
LIML	2.969(1.503)	(0.023, 5.916)	-	-	0.001
TSHT	0.861(0.245)	(0.380, 1.342)	{1}	{1}	0.999
CIIV*	2.635(1.974)	(-1.233, 6.504)	{2,4,5,6,7}	-	0.385
sisVIVE	0.819(-)	-	{1,2,4}	-	0.418
Post-Alasso	0.964(0.251)	(0.471, 1.457)	{1,2,4,5,6}	-	0.086
WIT	0.974(0.323)	(0.340, 1.609)	{1,2,4,6}	-	0.275

Note: CIIV* stands for CIIV method without first stage IVs selection because it reports that "Less than two IVs are individually relevant, treat all IVs as strong". Sargan test means p-value of Sargan test and selection of relevant IVs $\hat{\mathcal{S}}$ is only be implemented in TSHT and CIIV.

Empirical findings

- p-value of the Hausman test for endogeneity is 0.000181 using the proxy for trade as IV.
- LIML using all potential IVs (without distinguishing the invalid ones) likely overestimates the treatment effect. The 0.001 p-value of Sargan test strongly reject the null of all potential IVs are valid.
- $\boldsymbol{Z}_5$ should be a invalid IV:
  1. In view of marginal correlation in Fig. **??**, $\boldsymbol{Z}_5$ is nearly uncorelated to trade but significantly correlated with log(GDP).
  2. Concerning the Sargen Test, p-value of $0.086 < 0.1$ in Post-Alasso indicates $\boldsymbol{Z}_5$ is not very credible to be valid.
  3. In the economic perspective, more arable land generates higher crop yields and maintains a higher agriculture sector labor force, which directly affects GDP.
- Strong IVs based method fails.

Reference

Anderson, J. E. (1979). A theoretical foundation for the gravity equation. *Am. Econ. Rev.*, 69(1):106–116.

Fan, Q. and Zhong, W. (2018). Nonparametric additive instrumental variable estimator: A group shrinkage estimation perspective. *J. Bus. Econ. Statist.*, 36(3):388–399.

Ending

Thanks!