Evaluating machine learning instrumental variable methods to estimate conditional treatment effects in Mendelian randomization

Marc-André Legault[1,2,*], Jason Hartford[3], Michael Lu[1], Archer Y. Yang[5], Joëlle Pineau[1,2]

Introduction

- Instrumental variable (IV) estimation is a method to estimate the causal effect of an exposure on an outcome in the presence of unmeasured confounding

- Mendelian randomization (MR) is the use of genetic variants as instrumental variables [1]

- Here, we evaluate nonparametric and semi-parametric IV estimators in the context of MR using realistic simulation settings [2,3]

- We consider a real world application to sclerostin, the drug target of sclerostin inhibitors (*e.g.* romosozumab), used to treat osteoporosis and with possible cardiovascular effects [4]

Research question: Can nonparametric instrumental variable estimators be used for Mendelian randomization?

Methodological background

Causal assumptions

- Relevance
- Unconfoundedness of the IV
- Exclusion restriction

Modeling assumptions

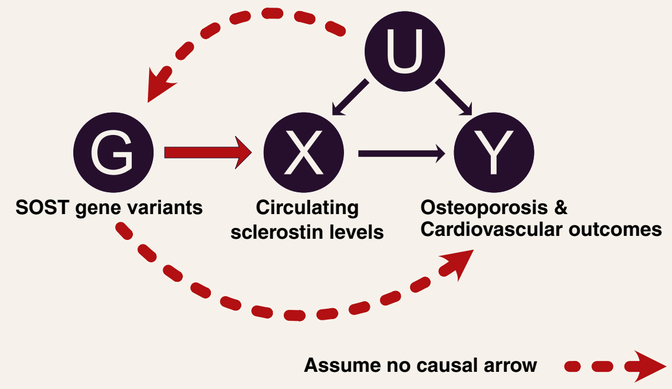
- For conventional methods: linearity & homogeneity

Estimators

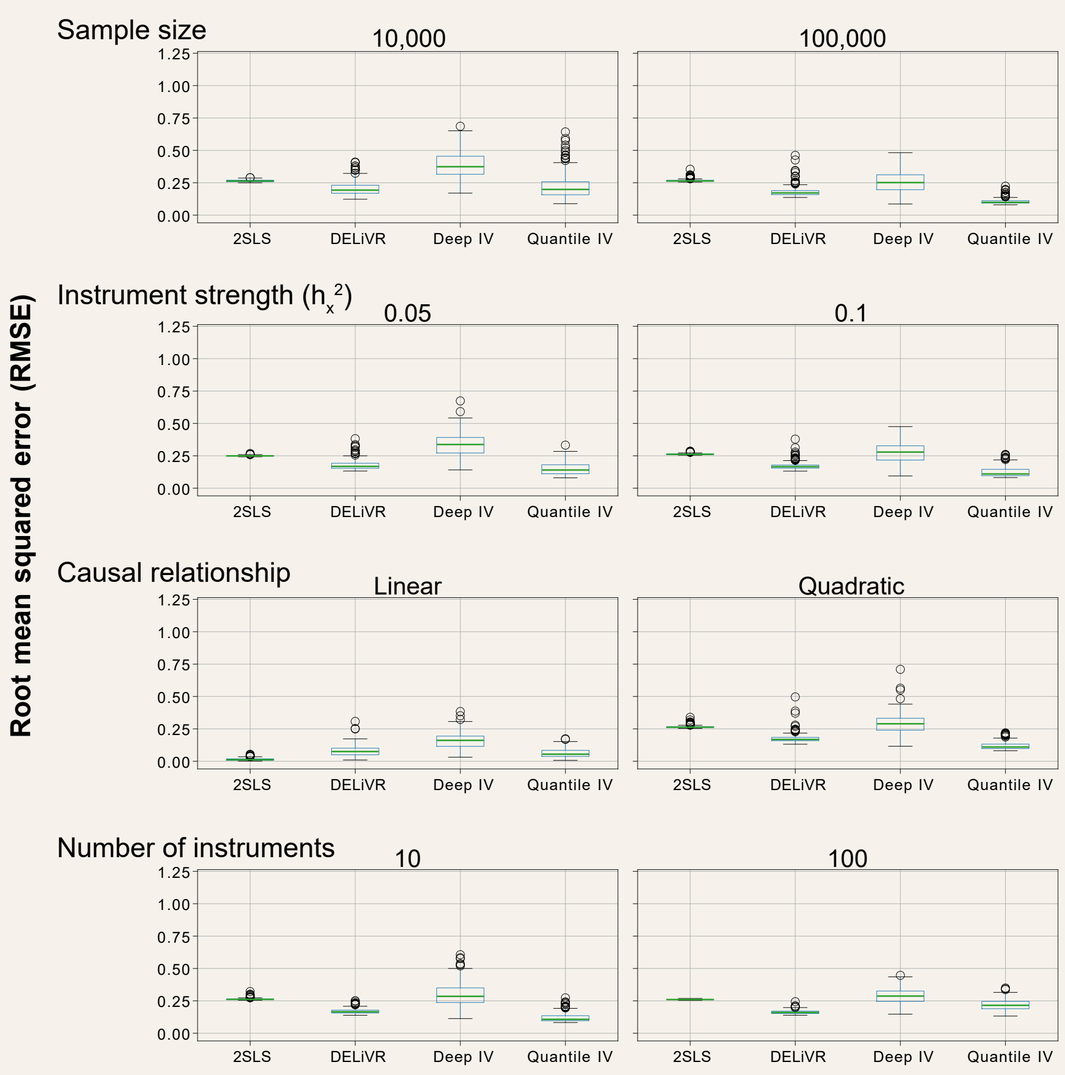
Parametric: 2 stage least squares:

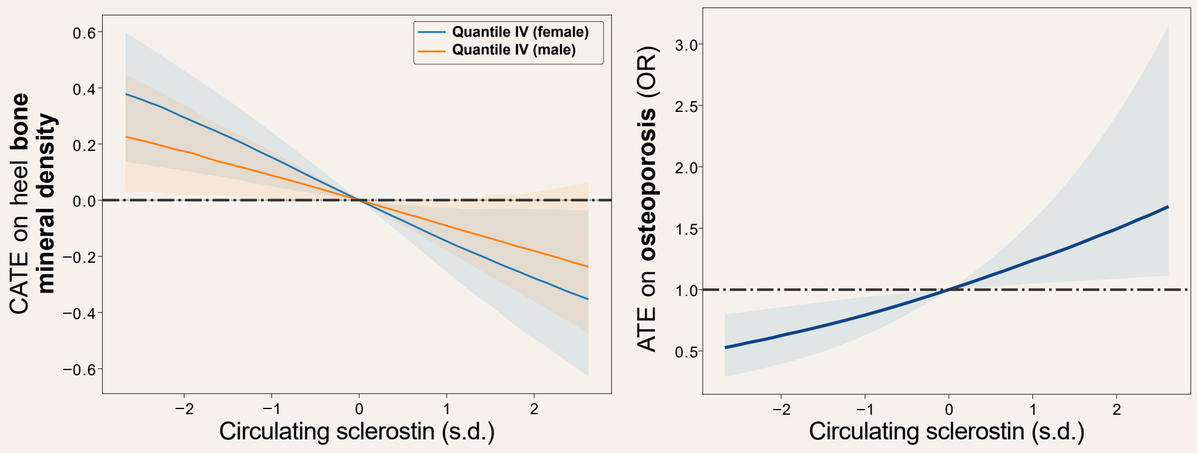
$$\mathbb{E}[X|G] = G\beta \qquad \hat{X} = G\hat{\beta} \qquad \mathbb{E}[Y|G] = \theta\hat{X}$$

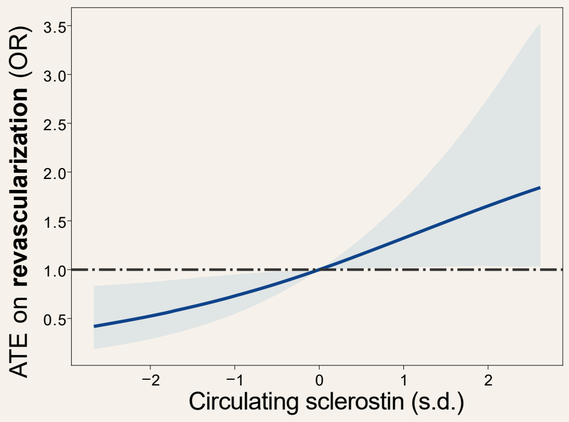
Nonparametric (*e.g.* DeepIV [3]): $\mathbb{E}[Y|G] = \int h(X)dF(X|G)$



Results







Top: Selected estimates from the real data application measuring the causal effect of varying circulating sclerostin levels about the mean in the UK Biobank. Conditional average treatment effect (CATE) estimation is straightforward in nonparametric IV. Shaded areas represent 90% bootstrap intervals (200 resamples).

Left: Average root mean squared error to the true structural equation between the empirical 2.5% and 97.5% of the exposure distribution across MR simulation scenarios.

Methods

Simulation study

- Draw independent genotypes with frequency from a Beta(1,3) distribution, sample effects following the baseline LDAK model and scale effects to achieve desired heritability (following [5])

- Confounding simulated implicitly by the correlation between the exposure and outcome (following [2])

- Compare 2SLS, Deep IV, DeLIVR and a new estimator based on DeepIV that avoids sampling using quantile regression in the first stage (Quantile IV)

Simulation parameter	Simulated Values
Structural equation	$0.4x$, $\mathbf{0.2(x-1)^2}$, $\max(x-2,0)$
Sample size	10 000, **50 000**, 100 000
Heritability of the exposure	**0.1**, 0.2, 0.5
Error correlation (confounding strength)	-0.6, **0.3**, 0.6
Number of independent instrumental variables	2, **10**, 100

Real data application

- Individual-level data from the UK Biobank was used for the analyses of bone mineral density, osteoporosis and cardiovascular outcomes.

- We use a split-sample approach based on 41,560 individuals with circulating sclerostin measurements (Olink) and the remaining UK Biobank participants (after genetic QC) for the outcome datasets

Conclusion

- We showed that our nonparametric IV estimator performs well across a wide range of genetically plausible simulations settings

- Nonparametric IV estimators can be used in a two-sample setting (but not with summary statistics)

- Bootstrapping can be used to quantify uncertainty

- Our estimates of the effect of a reduction of sclerostin are concordant with results from clinical trials showing that pharmacological sclerostin inhibition increases bone mineral density and protects against osteoporosis

- In the current study, sclerostin inhibition decreases the risk of revascularization procedures suggesting a possible atheroprotective effect

Affiliations

1. Department of Computer Science, McGill University, Montreal, Canada
2. Mila, Montreal, Canada
3. Recursion, Salt Lake City, United States of America
4. Department of Mathematics and Statistics, McGill University, Montreal, Canada

References

1. Didelez, V. & Sheehan, N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat. Methods Med. Res.* 16, 309–330 (2007).
2. He, R. et al. DeLIVR: a deep learning approach to IV regression for testing nonlinear causal effects in transcriptome-wide association studies. *Biostatistics* (2023)
3. Hartford, J., Lewis, G., Leyton-Brown, K. & Taddy, M. Deep IV: A Flexible Approach for Counterfactual Prediction. *ICML* (2017).
4. Golledge, J. & Thanigaimani, S. Role of Sclerostin in Cardiovascular Disease. Arteriosclerosis, *Thrombosis, and Vascular Biology* 42, e187–e202 (2022).
5. Sulc, J., Sjaarda, J. & Kutalik, Z. Polynomial Mendelian randomization reveals non-linear causal effects for obesity-related traits. *HGG Adv* 3, 100124 (2022).

✉ marc-andre.legault@mcgill.ca