When is TSLS actually LATE?

When is TSLS weakly causal?

Yichen Han

2025-02-07

yichen.han@campus.lmu.de

Seminar presentation based on Blandhol et al., 2022, revised Jan 14 2025

Outline

1.1 The Common TSLS Recipe

Selection on unobservables

We have an outcome variable Y, a discrete, ordered treatment $T \in \mathcal{T} = \{t_0, t_1, ..., t_J\}$, a discrete instrument $Z \in \mathcal{Z}$.

First Stage:

$$T = \alpha_1 + \beta_1 Z + v$$

Second Stage:

$$Y = \alpha_2 + \beta_{\text{tsls}} \hat{T} + \varepsilon$$

We assume:

- Relevance, i.e. $\text{Cov}(Z, T) \neq 0$,
- Exclusion, i.e. $\text{Cov}(Z, \varepsilon) = 0$.
- Monotonicity, i.e. no defiers.

$\beta_{\text{tsls}} = \frac{\text{Cov}(Y,Z)}{\text{Cov}(T,Z)}$ identifies LATE of compliers. There is nothing wrong with this classcial setup.

1.1 The Common TSLS Recipe

Selection on unobservables & observables

We have $Y, T \in \mathcal{T} = \{t_0, t_1, ..., t_J\}$, $Z \in \mathcal{Z}$, and an observed covariate matrix X.

First Stage:

$$T = \alpha_1 + \beta_1 Z + \gamma_1 X + \upsilon$$

Second Stage:

$$Y = \alpha_2 + \beta_{\text{tsls}} \hat{T} + \gamma_2 X + \varepsilon$$

We additionally assume **exogeneity (EX)**: conditional on X, the assignment of Z is independent of both the potential treatment and potential outcomes, i.e. $(\{T(z)\}_{z\in\mathcal{Z}}, \{Y(t)\}_{t\in\mathcal{T}}) \perp Z \mid X$.

Typically, β_{tsls} does not have a causal interpretation and is not LATE.

1.1 The Common TSLS Recipe

Angrist and Pischke, 2009:

"TSLS with covariates produces an average of covariate-specific LATEs." – under saturated covariates and instrument (Saturate and weight, SW).

IV Paper (n=122)	N	%
Used TSLS	112	92
TSLS + Covariates	99	81
TSLS + Covariates, as LATE	30	25

TSLS + Covariates (n=99)	N
Followed SW	1
Saturated in covariates	4
Not saturated in covariates	94

General critiques on IV-LATE: (Robins and Greenland, 1996; Heckman, 1997; Angrist and Imbens, 1999; Deaton, 2010; Imbens, 2010; Swanson and Hernan, 2014)

- complier-specific, misleading for population analysis;
- vulnerable to heterogeneous treatment effects;
- assumptions like EX are hard to empirically justify.

1.2 Motivation

Def. Saturate and weight (SW):

Given a treatment variable T, let X take values in $\mathcal{X} = \{x_1, ..., x_L\}$, and Z in $\mathcal{Z} = \{z_1, ..., z_K\}$. The TSLS model takes C as covariates and I as instrument, where:

- $C = [1, \mathbb{I}(X = x_l) : l = 2, ..., L]'$, i.e., one coefficient for each possible x on its entire support;
- $I = [\mathbb{I}(Z = z_k), \mathbb{I}(Z = z_k)\mathbb{I}(X = x_l) : l = 2, ..., L; k = 2, ..., K]'$, i.e. the instrument vector consists of each Z value and the interaction term between each specific z, x on their supports.

SW is not possible if any of the variables takes continuous values, and is rarely feasible for empirical studies that include many covariates.

SW may also incur many instruments bias due to the excessive instruments generated from $\mathbb{I}(Z = z_k)\mathbb{I}(X = x_l)$, biasing it towards the OLS estimator.

1.2 Motivation

Practical relevance:

- TSLS with covariates is popular, with β_{tsls} often interpreted as LATE of compliers;
- SW is a highly theoretical setup, and is not favoured by empirical studies;
- Without SW, does the LATE interpretation still apply?

Unsolved problem:

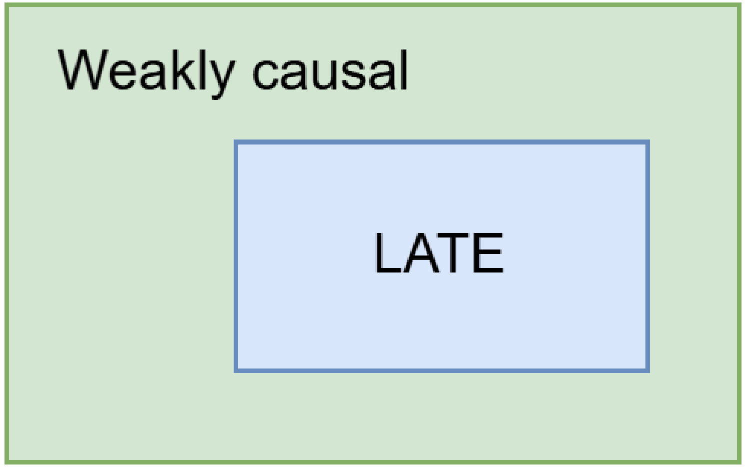
- Angrist and Krueger, 1999: [With IV estimates in models with unsaturated covariates], it seems reasonable to assume that some sort of approximate weighted average is being generated, but we are unaware of a precise causal interpretation that fits all cases.

Related effort:

- Słoczynski, 2020, 2024: focuses on incorrect monotonicity and thus the impact of negative weights.
- Evdokimov & Kolesár, 2018: on treatment effect heterogeneity, and sufficient conditions for LATE.
- Abadie, 2003: theoretical refinement under constant treatment effect assumption.

1.2 Motivation

$$\beta_{\text{tsls}} \notin$$



Outline

2 Binary T and Z

How education affects income? From Card (1995), adpated.

- $N = 3010$ individuals
- Y: continuous, log hourly wage
- T: binary, received education > 12 yrs
- Z: binary, presence of college nearby
- X: 4 binary covariates: work experience > 8 yrs, 3 indicators for race and location of residence.

We assume:

1. Exogeneity: $Z \perp (Y(0), Y(1), T(1), T(0)) \mid X$;
2. Monotonicity: $\mathbb{P}[T(1) \geq T(0)] = 1$ (no defiers).

2 Binary T and Z

Since T, Z, X are binary, we can do SW.

```r
# OLS
lm(Y ~ D + X)
# IV without saturation:
ivreg(Y ~ X + D | X + Z)
# IV with saturated covariates:
Z_int <- cbind(Z, Z * X1, Z * X2, Z * X3, Z * X4, Z * X1 * X2, Z * X1 * X3, Z * X1 * X4, Z * X2 * X3, Z * X2 * X4, Z * X3 * X4, Z * X1 * X2 * X3, Z * X1 * X2 * X4, Z * X1 * X3 * X4, Z * X2 * X3 * X4, Z * X1 * X2 * X3 * X4)
ivreg(Y ~ X1*X2*X3*X4 + D | X1*X2*X3*X4 + Z_int)
```

Method	Coefficient on D	SE
OLS	0.185 ***	0.016
TSLS, no covar.	1.317 ***	0.232
TSLS, unsaturated	0.778 *	0.353
TSLS, SW	0.412 *	0.179

RESET-test on `lm(Z~X)`: $p = 0.000$, i.e. no rich covariates.

Overestimated causal effect: $\frac{\beta_{\text{tsls}} - \beta_{\text{sw}}}{\beta_{\text{tsls}}} = 47\%$?

Overestimated selection bias: $\left|\frac{\beta_{\text{tsls}} - \beta_{\text{sw}}}{\beta_{\text{ols}} - \beta_{\text{sw}}}\right| = 161\%$?

Consider many instruments bias.

2 Binary T and Z

We divide the population into groups according to potential treatment:

$$G := (T(0), T(1)) = \begin{cases} (0,1) =: \mathrm{CP} \\ (1,1) =: \mathrm{AT} \in \mathcal{G} \\ (0,0) =: \mathrm{NT} \end{cases}$$

Denote the conditional average treatment effect as $\mathrm{CATE}(g,x) := \mathbb{E}\big[Y(1) - Y(0) \mid G = g, X = x\big]$.

We introduce one of the key findings, Prop.2, in a minimalist style.

Proposition 2. Assume potential outcome is linearly predicted by X.

Then, the β_{tsls} can be decomposed into the sum of weighted average of CATEs of CP, AT, and NT. That is:

$$\beta_{\text{tsls}} = \sum_{g \in \mathcal{G}} \mathbb{E}[\omega(g,x) \cdot \mathrm{CATE}(g,x)]$$

where for AT and NT, $\omega(g,x) = \mathbb{E}\big[\tilde{Z}|X\big]\mathbb{P}[G = g|X]\mathbb{E}\big[\tilde{Z}T\big]^{-1}$, while $\omega(\mathrm{CP}, x) > 0$.

2 Binary T and Z

- The signs of weights on AT and NT are determined by $\mathbb{E}\left[\tilde{Z}|X\right]$.
- $\tilde{Z}$ is the residuals from the regression $Z \sim X$.
- If and only if $\mathbb{E}\left[\tilde{Z}|X\right] = 0$, a condition defined as having **rich covariates**, β_{tsls} only reflects the positively weighted treatment effects on compliers, and is thus LATE.
- If without rich covariates, β_{tsls} reflects all three groups. As $\mathbb{E}\left[\tilde{Z}|X\right]$ is possible to take negative values for any data, AT and NT can be negatively weighted, thus biasing the direction of treatment effect interpretation.
- Negative weighting on some treatment groups is a violation to what the authors define as **weakly causal**.

Outline

3 What is Weak Causality?

Assume throughout that we have $T \in \{0, 1\}$, $Z \in \{0, 1\}$.

Def. Weakly causal (WC): an estimand β is WC iff :

$$\forall g \in \mathcal{G}, \text{and every } x \in \mathcal{X}, \text{if } \text{CATE}(g, x) \geq 0, \text{then } \beta \geq 0$$
$$\forall g \in \mathcal{G}, \text{and every } x \in \mathcal{X}, \text{if } \text{CATE}(g, x) \leq 0, \text{then } \beta \leq 0$$

Intuitively, if taking the treatment always gives higher potential outcome for every group, then the estimator should also be positive. A WC estimator is not systematically wrong in its direction.

Under this weak criterion, the trivial estimand $\beta = 0$ is also WC.

We want to further investigate some parametric or semi-parametric settings that translate into weak causality.

3 What is Weak Causality?

The decomposition in Prop.2 can be rewritten with a different definition of weights as:

Proposition 3.

$$\beta = \sum_{g,x} \overbrace{\omega_0(g,x)\mathbb{E}[Y(0)|g,x]}^{\text{baseline level w/o treatment}} + \sum_{g,x} \overbrace{\omega_1(g,x)\text{CATE}(g,x)}^{\text{diff. in outcome w. treatment}}$$

where $\omega_0(g,x) = \mathbb{E}\left[\tilde{Z}T\right]^{-1}\mathbb{E}\left[\tilde{Z}|g,x\right]\mathbb{P}(g,x)$, and $\omega_1(g,x) = \mathbb{E}\left[\tilde{Z}T\right]^{-1}\mathbb{E}\left[\mathbb{I}(T=1)\tilde{Z}|g,x\right]\mathbb{P}(g,x)$

Proposition 4. Given **BF**, **EX**, and suppose further **GR**, then β is WC iff

1. (Non-negative weights): $\forall g, x,\ \omega_1(g,x) \geq 0$;
2. (Baseline irrelevance): $\forall g, x,\ \omega_0(g,x) = 0$.

Intuitively,

- Each choice group is strictly non-negatively weighted;
- The baseline level of outcome where the treatment is absent does not flow into β.

Outline

4 When do we have WC?

For binary T and Z.

Assumption EX: $Z \perp (Y(0), Y(1), T(1), T(0)) \mid X$

Def. Rich covariates (RC): Let $\mathbb{L}[Z|X = x]$ be the population fitted values at $X = x$ from the linear regression $Z \sim X$. An IV specification has RC if:

$$\forall x \in \mathcal{X},\ \mathbb{L}[Z|X = x] = \mathbb{E}[Z|X = x] \Leftrightarrow \forall x \in \mathcal{X}, \mathbb{E}\left[\tilde{Z}|X\right] = 0$$

Saturation guarantees **RC**. If we have **RC** without saturating, it's called "correctly specified".

Assumption MN: Monotonicity. $\forall x \in \mathcal{X},\ \mathbb{P}[T(0) \leq T(1) \mid X = x] = 1$.

Intuitively, receiving the instrument does not decrease potential treatment, for all x.[1]

[1] Strong monotonicity. For results under weaker restrictions see Blandhol et al. 2022 (pre-revision).

4 When do we have WC?

Theorem 1. Suppose **EX** and **MN** are satisfied, then β_{tsls} is WC if and only if **RC**.

Intuitively, Th.1 states that without **RC**, then baseline-irrelevance in Prop.4 & 5 does not hold, and the estimator is not WC.

Recall baseline irrelevance, which requires:

$$\forall g, x,\ \omega_0(g, x) = \mathbb{E}\big[\tilde{Z}T\big]^{-1}\mathbb{E}\big[\tilde{Z}|G = g, X = x\big]\mathbb{P}(g, x) = 0$$

From **EX**, we have $\tilde{Z} \perp G \mid X$. Therefore, $\mathbb{E}\big[\tilde{Z}|G = g, X = x\big] = \mathbb{E}\big[\tilde{Z}|X = x\big]$. With $\mathbb{P}(g, x) > 0$, this holds iff $\mathbb{E}\big[\tilde{Z}|X = x\big] = 0$, i.e. **RC**.

The $\tilde{Z}$ problem stems from the practice of interpreting the linear TSLS estimator through yet a nonparametric setup. The linear-regression-based TSLS was meant to serve linear IV models.

4 When do we have WC?

If **MN** fails, **RC** does not imply an WC estimator[1]. We have to additionally restrict treatment effect homogeneity.

Assumption CLE: constant, linear effects. Order T into $\{t_0, t_1, ..., t_J\}$. There exists an $\delta \in \mathbb{R}$, such that $\mathbb{E}\left[Y(t_j) - Y(t_{j-1}) \middle| g, x\right] = \delta(t_j - t_{j-1})$ for every $j \geq 1, g \in \mathcal{G}, x \in \mathcal{X}$.

Intuitively, the mean difference in potential outcome is linearly proportional to the difference in treatment levels, for all groups and all x.

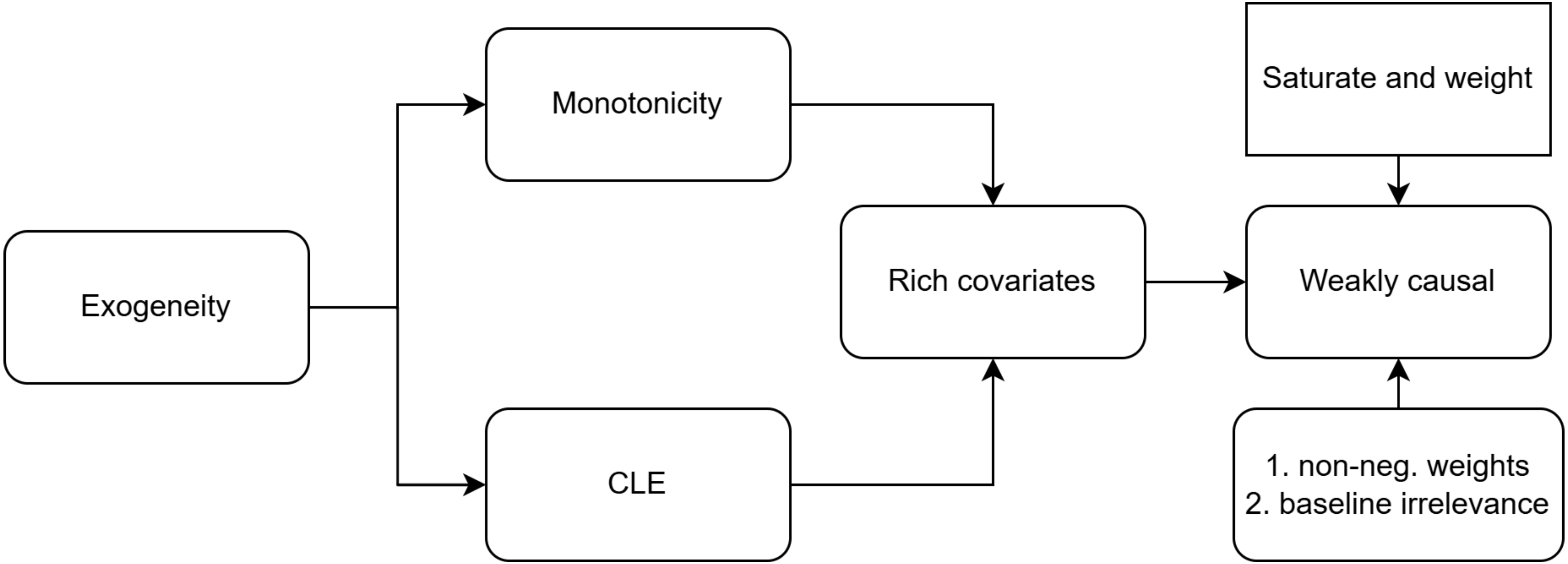
For binary T, Z, this simplifies to $\forall g, x, \text{CATE}(g, x) = \delta$.

Proposition 5. Suppose **EX** and **CLE** are satisfied, then β_{tsls} is WC iff **RC**.

[1]Suggested reading: Słoczynski 2024, When Should We (Not) Interpret Linear IV Estimands as LATE?

4 When do we have WC?

Wrap-up:



Outline

5.1 The New Recipe

1. If covariates are not essential for justifying instrument exogeneity, then do TSLS without covariates.
2. If covariates need to be included, think twice on including which covariates. The more included, the more likely that the specification is not rich in covariates.
3. Use RESET-test (Ramsey, 1969) `lmtest::resettest()` to examine covariate richness.
4. In general, the linear TSLS estimator can be complemented with:
   - DDML-PLIV, which produces $\beta_{\text{pliv}} \approx \beta_{\text{rich}}$.
   - If Z, T are binary, DDML-LATE identifies LATE of compliers.

5.2 Intuition

We have discussed that $\tilde{Z}$ became a problem because we tried to impose the linear IV model onto an unknown data structure which, in practice, is hardly linear.

Instead, we consider the following **partially linear IV model (PLIV)**:

$$Y = \beta_{\text{pliv}} T + f(X) + \varepsilon, \quad \mathbb{E}[\varepsilon | X, Z] = 0$$
$$Z = g(X) + v, \quad \mathbb{E}[v | X] = 0$$

That is, we only assume linearity on T, and try to approximate the non-linear functions f, g. A reasonable way is to do machine learning.

Problems: causal interpretation is vulnerable to overfitting, regularization bias and estimation error. **DDML** (Double/debiased machine learning) uses orthogonalization and data splitting to solve these.

5.2 Intuition

DDML also applies a group of base learners, e.g. OLS, LASSO, XGBoost, RF, and then outputs the β_{pliv} estimate from an ensemble model.

It can be proven that β_{pliv} approximates β_{rich} under correctly specified linear IV, without additional parametric assumptions.

If we have $T \in \{0, 1\}$ and $Z \in \{0, 1\}$, then a three-stage framework will approximate the LATE estimator for compliers β_{late}.

Implemented in the R-Package **ddml** (Chernozhukov et al., 2018; Ahrens et al., 2024).

$\boldsymbol{\beta_{\text{rich}} \neq \beta_{\text{late}}}$**!** Under binary T, Z,
- β_{late} is the simple mean effect for covariate-specific complier groups, thus LATE.
- β_{rich} imposes weights on subgroups dependent on $\mathbb{V}[Z|X]$ – preference for subgroups with more variance in Z.

Outline

6.1 Card (1995)

We extend the previous adapted Card (1995) experiment.

```r
library(ddml)
set.seed(123)
# Estimate the PLIV using short-stacking with base learners ols, rlasso, random forest, and xgboost.
learners_multiple <- list(list(fun = ols), list(fun = mdl_glmnet), list(fun = mdl_ranger), list(fun = mdl_xgboost))
pliv_fit_short <- ddml_pliv(Y, D, Z, X,
  learners = learners_multiple,
  ensemble_type = c('nnls', 'singlebest', 'average'),
  shortstack = TRUE,
  sample_folds = 10, cv_folds = 10)
```

Method	Coefficient on D	SE
OLS	0.185	0.016
TSLS, no covar.	1.317	0.232
TSLS, unsaturated	0.778	0.353
TSLS, SW	0.412	0.179
TSLS, sat. in X	0.913	0.439
DDML-PLIV	0.912	0.432
DDML-LATE	0.662	0.325

- Speculated many instruments bias: 38%
- Overestimated causal effect: $\frac{\beta_{\text{tsls}}-\beta_{\text{pliv}}}{\beta_{\text{tsls}}} = -17\%$
- Overestimated selection bias: $\left|\frac{\beta_{\text{tsls}}-\beta_{\text{pliv}}}{\beta_{\text{ols}}-\beta_{\text{pliv}}}\right| = 18\%$

6.1 Card (1995)

The original Card (1995) data:

- $N = 3010$ individuals
- Y: cont. log hourly wage
- T: cont. years of education
- Z: binary, presence of college nearby
- X: cont. years of experience, and indicators on race, location of residence, etc. (14 co-variates)
- RESET-test on `lm(Z~X)`: $p = 0.000$

Method	Coefficient on D	SE
OLS	0.075	0.003
TSLS, no covar.	0.188	0.026
TSLS, unsaturated	0.132	0.054
TSLS, SW	-	-
TSLS, sat. in X	-	-
DDML-PLIV	0.138	0.050
DDML-LATE	0.068	0.043

- Overestimated causal effect: -4.5%
- Overestimated selection bias: 9.5%
- Relatively low bias among other works.

6.2 Nunn and Wantchekon (2011)

Effect of the slave trade on modern day level of trust in Africa. The data[1]:

- $N = 16679$ individuals
- Y: cat. level of trust in neighbors
- T: cont. log of total slave exports
- Z: cont. distance from the nearest coast
- X: 93 covariates.
- RESET-test on lm(Z~X): $p = 0.000$

Method	Coefficient on D	SE
OLS	−0.203	0.033
TSLS, no covar.	−0.190	0.111
TSLS, unsaturated	−0.271	0.088
TSLS, SW	-	-
TSLS, sat. in X	-	-
DDML-PLIV	−0.071	0.091
DDML-LATE	-	-

- Overestimated causal effect: 73.8%
- Overestimated selection bias: 151.5%
- The authors' key finding was invalid!
- Cited by 2800+.

[1]Result of this table is currently cited from Table 3(2), Blandhol et al. 2025.

6.3 MacDonald et al. (2018)

Increased human-forest interaction = more Lyme disease incidence? The data[1]:

- $N = 514$ counties, USA
- Y: cont. log Lyme disease incidence
- T: cont. population% living near forest
- Z: cont, land use regulation score
- X: year, forest%, forest edge density, average forest patch area in county (4 cont.)
- Binary: T>median, Z>0.

	Original	**Binary**
OLS	0.008 (0.004)	0.332 (0.262)
TSLS, no covar.	0.037 (0.009)	2.995 (0.967)
TSLS, unsaturated	0.037 (0.014)	3.473 (1.910)
DDML-PLIV	0.030 (0.009)	2.872 (1.522)
DDML-LATE	-	0.922 (0.693)
RESET p-Value	0.579	0.0001

- Careful theoretical justification
- Ignored saturation, but had RC (by chance?)
- Interpreted as in simple linear relationship, not directly mentioning LATE or causality.
- Systematic bias of DDML is visible.

[1]Only complete cases, used cluster-robust standard errors with contiguous counties as clusters.

Outline

7 Discussion

Contributions

- Extended Angrist and Krueger 1999, by providing a precise interpretation for the TSLS estimate with covariates that is generally applicable.
- Proved that TSLS with covariates, if not under strong assumptions, cannot provide weakly causal interpretation, and thus cannot identify LATE.
- Extended previous efforts from, e.g. Słoczynski and Kolesár, by combining covariate richness, treatment effect heterogeneity, and monotonicity under a more holistic investigation with more general setups.
- For empirical researchers, it showed TSLS with covariates is a common practice, urges caution when doing so, provided a guided recipe on choosing IV specifications, and also discussed a few alternative methods to linear IV.

7 Discussion

Limitations

- It is clear that LATE $\Rightarrow$ WC, but how WC relates to LATE needs to be specified. At least we know, empirically, $\beta_{\text{rich}} \neq \beta_{\text{late}}$. The weights are empirically inaccessible.
- Similarly, the estimator DDML-PLIV is generally hard to interpret. Only in binary cases, we get DDML-LATE for compliers. Further research is needed.

Notes

- Further theoretical and empircial investigation could be done on analogous methods, e.g. 2SRI. Consider a more thorough survey on previous empirical results in diverse fields.
- Call for good practice: Discussed results from 13 papers, selected based on publicly available data and reproducible results, out of 99 surveyed in total. Itself does not publish code.
- My codes are available on GitHub: https://github.com/Yc-Han/Causal_TSLS.

Outline

References

Blandhol, C., Bonney, J., Mogstad, M., & Torgovitsky, A. (2022). When is TSLS actually late? (No. w29709). Cambridge, MA: National Bureau of Economic Research.

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. Journal of econometrics, 113(2), 231-263.

Ahrens, A., Hansen, C., Schaffer, M., & Wiemann, T. (2024). ddml: Double/Debiased Machine Learning. R package version 0.3.0, https://CRAN.R-project.org/package=ddml.

Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in labor economics. In Handbook of labor economics (Vol. 3, pp. 1277-1366). Elsevier.

Angrist, J. D., & Pischke, J. S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton university press.

Angrist, J. D., Imbens, G. W., & Krueger, A. B. (1999). Jackknife instrumental variables estimation. Journal of Applied Econometrics, 14(1), 57-67.

Card, D. (1995). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp, edited by L. N. Christofides, K. E. Grant, and R. Swidinsky, Toronto: University of Toronto Press, 201-222.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Deaton, A. (2010). Instruments, randomization, and learning about development. Journal of economic literature, 48(2), 424-455.

Evdokimov, K. S., & Kolesár, M. (2018). Inference in Instrumental Variables Analysis with Heterogeneous Treatment Effects.

Heckman, J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. Journal of human resources, 441-462.

Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). Journal of Economic literature, 48(2), 399-423.

Kolesár, M. (2013). Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity. Working paper.

MacDonald, A. J., Larsen, A. E., & Plantinga, A. J. (2019). Missing the people for the trees: Identifying coupled natural–human system feedbacks driving the ecology of Lyme disease. Journal of Applied Ecology, 56(2), 354-364.

Nunn, N., & Wantchekon, L. (2011). The slave trade and the origins of mistrust in Africa. American economic review, 101(7), 3221-3252.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology, 31(2), 350-371.

Robins, J. M., & Greenland, S. (1996). Identification of causal effects using instrumental variables: comment. Journal of the American Statistical Association, 91(434), 456-458.

Słoczyński, T. (2020). When should we (not) interpret linear iv estimands as late?. arXiv preprint arXiv:2011.06695.

Swanson, S. A., & Hernán, M. A. (2014). Think globally, act globally: an epidemiologist's perspective on instrumental variable estimation. Statistical science: a review journal of the Institute of Mathematical Statistics, 29(3), 371.