# When is TSLS actually LATE? — Understanding the Misspecification Bias in TSLS with Covariates

Department of Statistics
Ludwig-Maximilians-Universität München

**Yichen Han**

Munich, March 10th, 2025

Seminar Term Paper for Seminar on Causal Inference, WT 2024/25
Supervised by Prof. Daniel Wilhelm

## Abstract

When instrumental variables (IV) are used to estimate causal effects, two-stage least squares (TSLS) is a popular choice. A survey of empirical research between 2018 and 2020 shows that covariates are frequently included in TSLS, with the resulting estimand often interpreted as the local average treatment effect (LATE) among compliers. However, recent work (Blandhol et al., 2025) shows that unless the covariates flexibly partial out the instrument—so-called "rich covariates"—this interpretation can fail, yielding a misspecified TSLS estimand with potentially negative weights on certain subpopulations. Under strong monotonicity, the estimand remains a positively weighted average of subpopulation treatment effects *if and only if* the specification includes rich covariates; otherwise, the TSLS coefficient may not even have the correct sign.

This paper synthesizes these findings and illustrates their empirical relevance. After reviewing prior work on TSLS specifications, we restate the sufficiency and necessity conditions for an estimand to be "weakly causal". We then present a simulation framework showing how the misspecification bias can contaminate the estimated causal effect, followed by two empirical examples. Next, we survey subsequent empirical and theoretical studies to examine the influence of the findings among practitioners. Finally, we discuss the contribution and limitations of the paper.

# Contents

# 1 Introduction

In econometrics, using instrumental variables (IV) to account for endogeneity has become a common approach. A regressor is endogenous when it is correlated with unobserved factors in the population. For example, in an ordinary least squares (OLS) regression $Y = \beta_0 + \beta_1 D + u$, if $\mathrm{Cov}(D, u) \neq 0$, $D$ is an endogenous regressor. In this case, the OLS estimand $\hat{\beta}_{\text{ols}}$ is biased. Introducing an instrumental variable $Z$ that is relevant and satisfies the exclusion restriction, i.e. $\mathrm{Cov}(Z, D) \neq 0$ and $\mathrm{Cov}(Z, u) = 0$, can isolate the exogenous variation and help identify the causal effect of $D$. Typically, monotonicity is also assumed. That is, increasing the instrument does not decrease the treatment.

A straightforward IV approach is two-stage least squares (TSLS). In its simplest form, where we assume the assignment of the treatment is based on unobservable factors, the two stages are given by:

- First Stage: $D = \alpha_0 + \alpha_1 Z + v$
- Second Stage: $Y = \beta_0 + \beta_{\text{tsls}} \hat{D} + \varepsilon$, where $\hat{D}$ is the fitted value from the first stage regression.

Since the work of Angrist and Imbens (1995), the estimand $\hat{\beta}_{\text{tsls}}$ is, under monotonicity, commonly referred to and interpreted as the local average treatment effect (LATE) on compliers, i.e. the subpopulation whose potential treatment changes in the direction of the instrument if the instrument is assigned differently.

However, in most empirical studies, it is natural to include and account for a set of observed covariates $X$ that may as well affect the assignment and outcome. That is,

- First Stage: $D = \alpha_0 + \alpha_1 Z + \gamma X + v$
- Second Stage: $Y = \beta_0 + \beta_{\text{tsls}} \hat{D} + \eta X + \varepsilon$

Recently, Blandhol et al. (2022) (hereafter B22) showed that, while interpreting the TSLS estimand with covariates as the LATE of compliers is a common approach following Angrist and Pischke (2009), the estimand is *not* a positively weighted average of treatment effects among compliers unless additional assumptions are satisfied. Therefore, the resulting estimand may not have a correct causal interpretation. In practice, the extent to which negative weights impact interpretation varies. However, given the policy relevance of economic studies, understanding and addressing this issue is beneficial for future practice.

This paper is mainly based on the revised version, Blandhol et al. (2025) (hereafter B25) and aims to summarize its main methodological message while extending it to additional empirical studies. Section 2 provides a brief summary of related theoretical and empirical works on TSLS prior to the publication of B22. Section 3 introduces its main result. Section 4 presents simulation and empirical examples. Section 5 surveys research articles published after the study. Finally, Section 6 summarizes the paper and discusses its limitations.

## 2 Literature Review

Prior to B22, including covariates in a TSLS model was a common practice among practitioners. It was only around 2020 that researchers recognized multiple restrictive, parametric assumptions to be *necessary* for a correct causal interpretation of the estimand. Meanwhile, several alternative or complementary methods were developed to enable more flexible and efficient treatment effect estimation.

Angrist and Pischke (2009) referred to the TSLS estimand with covariates as *an average of covariate-specific LATEs* under the so-called “saturate and weight” specification (SW, see Section 3.2), which additionally accounts for the full support of each of the covariates and the instrument and their interactions in both stages of TSLS. However, SW is not commonly applied by empirical studies.

B22 surveyed 122 empirical research articles published on five reputable journals between 2000 and 2018 involving instrumental variables (Table 1, adapted from Table 2, B22). Of these, 99 papers employed TSLS with covariates in both stages, suggesting the popularity of the approach. Among them, only one paper followed the SW specification, 4 saturated in covariates, and 94 did not include any model with saturation. In 30 of the 99 papers, the TSLS estimand was interpreted as LATE. This raises the concern that without SW as premise, the LATE interpretation may not apply.

However, SW is restricted to covariates with finite support, thus not applicable to datasets with continuous variables. Additionally, it could induce many-instrument bias, potentially biasing the estimand toward OLS (Chamberlain and Imbens, 2004). Consequently, further research on the necessary and/or sufficient conditions for interpreting the TSLS estimand as LATE ensued. The following two papers worked closely on this topic.

Evdokimov and Kolesár (2018) first pointed out that, under monotonicity and treatment effect heterogeneity, the conditional (on covariates and instruments) treatment effect differs from the unconditional average. The conditional estimand involves covariate-specifc

**IV Paper (n=122)**

| | N | % |
|---|---|---|
| Used TSLS | 112 | 92 |
| TSLS + Covariates | 99 | 81 |

**TSLS + Covariates (n=99)**

| | N | % |
|---|---|---|
| Saturated in covariates and instrument (SW) | 1 | 1 |
| Saturated in covariates | 4 | 4 |
| Not saturated | 94 | 95 |
| Interpreted the estimand as LATE | 30 | 30 |

Table 1: Overview of IV Papers and TSLS Usage, 2000-2018

weights and exhibits lower variability. They derived a decomposition of the estimand (Lemma 4.1) and proved that if all covariates included are group dummy indicators (a special case of *rich covariates*, see Section 3.2 ahead), then the weights are all positive. However, necessary conditions for the weights to be non-negative were not covered.

Słoczyński (2024) focused on the necessity of monotonicity. Specifically, assuming rich covariates, weak monotonicity, i.e. the potential treatment is monotonic in the instrument conditional on covariates, does not guarantee non-negative weights in the TSLS estimand unless the first stage regression includes interactions between the instrument and covariates. Under a special case, the study also showed that even when rich covariates and monotonicity are both satisfied, the non-negatively weighted average of LATEs can still substantially deviate from the unconditional average treatment effect, often viewed as the policy-relevant quantity for empirical researchers.

B25 extends the discussion with a general nonparametric setup and investigates both sufficient and necessary conditions for the estimand to be a convex aggregate of LATEs. It mainly focuses on the necessity of rich covariates, adopting the strong monotonicity of Słoczyński (2024), and provides practical suggestions for empirical studies to address the negative-weighting problem. Compared to B22, B25 is more practice-oriented and presents a simplified discussion of monotonicity.

Since TSLS implicitly assumes a linear relationship analogous to OLS, several less restrictive methods were developed to estimate the treatment effect using IV. For example, Abadie (2003) developed a $\kappa$-weighted regression strategy for binary instrument and treatment that allows for nonparametric estimation in the weights, and the estimand can be interpreted as the average causal response among compliers. However, its Proposition 5.1 and B25 made clear that unless the rich covariates assumption is satisfied, the resulting estimand is not equivalent to that of TSLS.

Recently, machine learning inspired strategies have been introduced to handle more complex situations. Athey et al. (2019) developed a versatile generalized random forest framework based on regularization, which can be adapted to perform IV regression. Chernozhukov et al. (2018) proposed double/debiased machine learning (DDML) based orthogonalization and ensemble learning. Using DDML, the partial linear IV regression framework (PLIV) is, according to B25, a good alternative to TSLS when key assumptions are difficult to justify. Details appear in Section 3.3.

# 3 Main Result

In this section, we present the main findings of B25, which demonstrate that strong monotonicity and "rich covariates" are *necessary* for TSLS with covariates to identify even a weaker notion of causality—termed "weakly causal". Each finding is followed by an intuitive sketch of the proof. By decomposing the TSLS estimand into covariate-specific weights and subpopulation treatment effects, B25 clarifies why misspecification of the model with covariates can lead to erroneous causal interpretations. The paper also offers practical guidance for empirical studies employing TSLS, illustrating how to mitigate negative weighting and discusses alternatives to TSLS.

Figure 1 summarizes these key ideas.

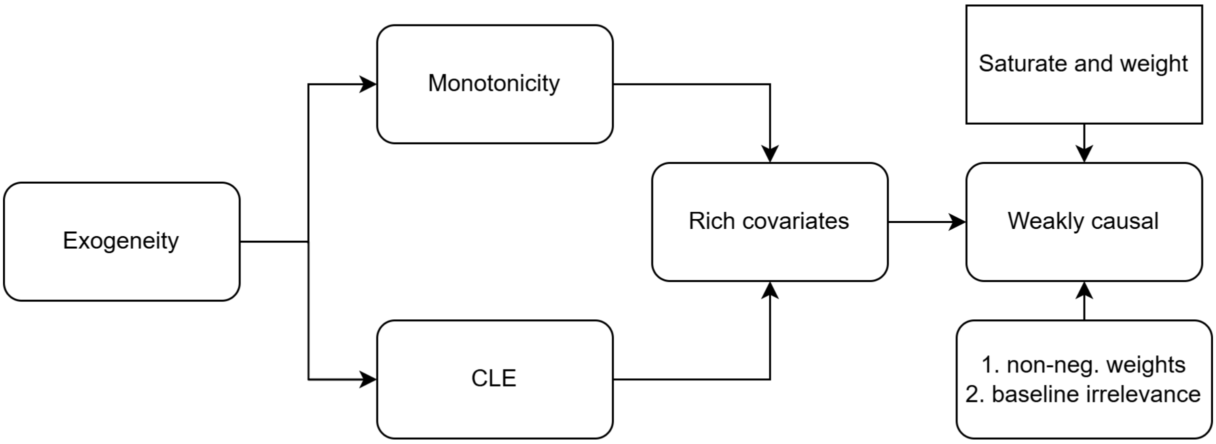

Figure 1: Summary of the main idea

## 3.1 Notation

In the following, we consider the nonparametric instrumental variable model.

Suppose we are interested in estimating the causal effect of a treatment $D$ on an outcome $Y$. We observe a set of covariates $X \in \mathcal{X}$, and an instrumental variable $Z$. For simplicity, we assume that both the instrument $Z$ and the treatment $D$ are discrete with finite support. Specifically, let $Z \in \mathcal{Z} = \{z_0, z_1, ..., z_K\}$, and $D$ is ordered such that $D \in \mathcal{D} = \{d_0, d_1, ..., d_J : d_0 \leq d_1 \leq ... \leq d_J\}$.

Throughout, we assume:

**Assumption 1.** Let $u$ denote the error term in the OLS regression $Y = \beta_0 + \beta_1 D + u$. A valid instrument $Z$ satisfies: (i) Exclusion: $\text{Cov}(Z, u) = 0$, and (ii) Relevance: $\text{Cov}(Z, D) \neq 0$.

Define the potential treatment given $Z = z$ as $D(z)$, and the potential outcome given $D = d$ as $Y(d)$. We omit the individual index $i$ for simplicity. The observed treatment and outcome relate to these potential quantities as follows:

$$D = \sum_{z \in \mathcal{Z}} \mathbb{I}(Z = z) D(z) \quad \text{and} \quad Y = \sum_{D \in \mathcal{D}} \mathbb{I}(D = d) Y(d)$$

The population can be partitioned into **subpopulation types** (or choice groups) based on their potential treatment. The choice group is defined by:

$$G := (D(z_0), D(z_1), ..., D(z_K)) \in \mathcal{G}$$

If $D$ and $Z$ are both binary, these subgroups are termed compliers (CP), never takers (NT), always takers (AT), and defiers (DF):

$$G := (D(0), D(1)) = \begin{cases} (0,1) & =: \ CP, \\ (1,1) & =: \ AT, \\ (0,0) & =: \ NT, \\ (1,0) & =: \ DF \end{cases}$$

Another necessary assumption for the validity of IV analysis is:

**Assumption 2.** (Exogeneity) $(G, \{Y(d)\}_{d \in \mathcal{D}}) \perp\!\!\!\perp Z \,|X$.

That is, conditional on $X$, the assignment of $Z$ is independent of both the potential treatment and potential outcome. This assumption is frequently involved in the proofs of the following propositions.

In the subsequent discussion, Assumptions 1 and 2 are maintained and thus not restated explicitly.

## 3.2 Theory

The TSLS estimand for a single treatment and a single instrument is given by:

$$\beta_{\text{tsls}} = \frac{\mathbb{E}[\tilde{Z}Y]}{\mathbb{E}[\tilde{Z}D]} = \mathbb{E}\left[\left(\frac{\tilde{Z}}{\mathbb{E}[\tilde{Z}D]}\right)Y\right] \tag{1}$$

We are interested in investigating whether $\beta_{\text{tsls}}$ produces a positively weighted average of covariate-specific LATEs, as Angrist and Pischke (2009) suggested, under a general model specification.

As LATE is clearly defined on compliers for **binary** $D$ and $Z$, we first consider this special case to showcase the problem that B25 tries to clarify.

Let $D, Z \in \{0,1\}$, we adopt the usual assumption that there are **no defiers**, i.e. $\mathbb{P}[D(1) \geq D(0)] = 1$. A natural motivation is to decompose the $\beta_{\text{tsls}}$ into a sum of treatment effects among the choice groups CP, AT, and NT. Ideally, the effects of AT and NT cancel out, while CP has non-negative weights. Therefore, we let

$$\text{CATE}(g, x) := \mathbb{E}[Y(1) - Y(0)|G = g, X = x]$$

denote the *conditional average treatment effect* for the subgroup $g \in \{AT, NT, CP\}$.

**Proposition 1.** (Prop.2, B25 pp.8) Suppose that $\mathbb{E}[Y(t)|X] = \eta_t' X$ for some parameters $\eta_t$, $t \in \{0,1\}$. Then for any real number $\epsilon$,

$$\begin{aligned}
\beta_{\text{tsls}} &= \sum_{g\in\mathcal{G}} \mathbb{E}\big[\omega_\epsilon(g, X)\text{CATE}(g, X)\big], \\
\text{where}\quad \omega_\epsilon(\text{CP}, X) &:= \Big(\epsilon\ \mathbb{E}[\tilde{Z}|X] + \mathbb{L}[Z|X](1 - \mathbb{E}[Z|X])\Big)\,\mathbb{P}(G = \text{CP}|X)\mathbb{E}[\tilde{Z}D]^{-1}, \\
\omega_\epsilon(\text{AT}, X) &:= \epsilon\ \mathbb{E}[\tilde{Z}|X]\mathbb{P}(G = \text{AT}|X)\mathbb{E}[\tilde{Z}D]^{-1}, \\
\omega_\epsilon(\text{NT}, X) &:= (\epsilon - 1)\mathbb{E}[\tilde{Z}|X]\mathbb{P}(G = \text{NT}|X)\mathbb{E}[\tilde{Z}D]^{-1}, \\
\mathbb{L}[Z|X] &:= X'\mathbb{E}[XX']^{-1}\mathbb{E}[X|Z], \\
\text{and}\quad \tilde{Z} &:= Z - \mathbb{L}[Z|X] \text{ are the residuals from a regression of } Z \text{ on } X.
\end{aligned}$$

According to Prop. 1, the signs of weights on AT and NT are determined by $\mathbb{E}[\tilde{Z}|X]$. If and only if $\forall x \in \mathcal{X}$, $\mathbb{E}[\tilde{Z}|X = x] = 0$, $\beta_{\text{tsls}}$ reflects the positively weighted treatment effects on compliers, and is thus LATE. As $\mathbb{E}[\tilde{Z}|X]$ is possible to take negative values for any data, AT and NT can be negatively weighted, thus biasing the direction of treatment effect interpretation. Numerically, the negative weights can flip the sign of the aggregate estimand, thus a violation to what B25 terms as **weakly causal**.

**Definition 1. Group treatment responses (GTRs)**:

$$\mu_j(g,x) := \mathbb{E}[Y(d_j)|G=g, X=x], \text{ and } \mu := \{\mu_j(g,x) : j = 0,1,...,J;\ g \in \mathcal{G};\ x \in \mathcal{X}\}$$

**Definition 2. Weakly causal (WC)**: an estimand $\beta$ is weakly causal if both of the following are true:

$$\text{If } \forall j \geq 1, g \in \mathcal{G}, x \in \mathcal{X},\ \mu_j(g,x) - \mu_{j-1}(g,x) \geq 0, \text{ then } \beta \geq 0;$$
$$\text{If } \forall j \geq 1, g \in \mathcal{G}, x \in \mathcal{X},\ \mu_j(g,x) - \mu_{j-1}(g,x) \leq 0, \text{ then } \beta \leq 0.$$

Intuitively, WC is a weak restriction– if taking a higher dose of the treatment always results in higher potential outcome for every group, then the estimand should be non-negative. A WC estimand is either trivial ($\beta = 0$) or has a correct sign.

The decomposition in Prop. 1 can be generalized. For the following, let $D \in \mathcal{D},\ Z \in \mathcal{Z}$.

**Proposition 2.** (Prop.3, B25 pp.15)

$$\beta = \sum_{g,x} \omega_0(g,x)\mu_0(g,x) + \sum_{g,x}\sum_{j=1}^{J} \omega_j(g,x)\left(\mu_j(g,x) - \mu_{j-1}(g,x)\right) \tag{2}$$

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

The proof of Prop. 2 mainly involves rewriting Equation 1 into a sum of conditional expectations over $g, x$, weighted by $\mathbb{P}(g,x)$. We refer to B25 for further details. In Equation 2, the first term involving $\mu_0(g,x)$ demonstrates the *level*, or the baseline outcome without receiving the treatment. When estimating the causal effect of the treatment, logically, $\omega_0(g,x)$ should always be 0. The second term is a weighted sum of subgroup- and covariate-specifc treatment effects. Reasonably, we expect $\omega_j(g,x)$ to be non-negative for all $j \geq 1$. Letting $\omega_0(g,x) \neq 0$ and/or some of the $\omega_j(g,x) < 0$, the $\beta$ can, in some cases depending on the data, have a different sign than $\mu_j(g,x) - \mu_{j-1}(g,x)$. This argument leads to the following:

**Proposition 3.** (Prop.4, B25 pp.15) $\beta$ is WC if and only if:

1. Non-negative weights: $\forall g, x,$ and $j \geq 1,\ \omega_j(g,x) \geq 0$, and
2. Level independence: $\forall g, x,\ \omega_0(g,x) = 0$.

According to Prop. 2, if $\omega_0(g,x) = 0$, we have $\forall g, x,\ \mathbb{E}[\tilde{Z}|g,x] = 0$. Under Assumption 2, $\mathbb{E}[\tilde{Z}|g,x] = \mathbb{E}[\tilde{Z}|X=x]$ since $\tilde{Z} \perp\!\!\!\perp G|X$, which is also required by Prop. 1. This argument *necessitates* the restriction that we, in the following, define as **rich covariates**.

**Definition 3. Rich covariates (RC)**: An IV specification has rich covariates if

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

Intuitively, $\mathbb{E}[\tilde{Z}|X]$ is of concern because we included $X$ linearly in both stages of TSLS, while not assuming the linearity under the nonparametric IV setup. To avoid this, $X$ should be included flexibly as, e.g. $C = f(X)$ so that $\mathbb{E}[Z|X] = \mathbb{L}[Z|C]$. The **SW specification** from Angrist and Pischke (2009) is a special case where RC is guaranteed. Furthermore, it is proved that SW satisfies both conditions in Prop. 3 and thus leads to a WC estimand.

**Definition 4. Saturate and weight (SW)**: Given a treatment variable $D$, let $X$ take values in $\mathcal{X} = \{x_1, \ldots, x_L\}$, and $Z$ in $\mathcal{Z} = \{z_1, \ldots, z_K\}$. The TSLS model takes $C$ as covariates and $I$ as instrument, where:

- $C = [1, \mathbb{I}(X = x_l) : l = 2, \ldots, L]'$;
- $I = [\mathbb{I}(Z = z_k), \mathbb{I}(Z = z_k)\mathbb{I}(X = x_l) : l = 2, \ldots, L; k = 2, \ldots, K]'$.

Sometimes we omit $I$ and only include $C$ and $Z$, which is called *saturated in X*. Intuitively, SW requires expanding the covariates into a matrix of dummy variables for all possible combinations over the entire support. For $n$ binary covariates, a saturated $C$ contains $2^n - 1$ dummies. The saturated $I$ additionally contains the interaction between every $z_k$ and $x_l$. Saturating is not possible if any of the $X$ is continuous, and is barely feasible if $X$ is high dimensional. Furthermore, as is discussed in B22, SW may induce the many instruments bias and requires a different estimation strategy, such as IJIVE or UJIVE.

For $\beta_{\text{tsls}}$ to be WC, we also need to account for the non-negative weights according to Prop. 3. First, we suppose that treatment effects are homogeneous, constant, and linear.

**Assumption 3.** (Constant, linear effects) There exists a constant $\Delta$ such that $\mu_j(g, x) - \mu_{j-1}(g, x) = \Delta(d_j - d_{j-1})$ for every $j \geq 1,\ g \in \mathcal{G},\ x \in \mathcal{X}$.

**Proposition 4.** (Prop.5, B25 pp.19) Suppose that Assumption 3 is satisfied, then $\beta_{\text{tsls}}$ is WC if and only if the IV specification has RC.

Under treatment effect heterogeneity, both B22 and Słoczyński (2024) have proven that the strong monotonicity is a necessary condition for WC.

**Assumption 4.** (Strong monotonicity) Label the values of $Z$ in increasing order as $z_0 \leq z_1 \leq ... \leq z_K$. Then

$$\forall x \in \mathcal{X},\ \mathbb{P}[D(z_0) \leq D(z_1) \leq ... \leq D(z_K)|X = x] = 1$$

Assumption 4 is "strong" in the sense that the direction of monotonicity is identical for all $x$, while weaker ones allow the direction to change depending on $X$, i.e. for some $X = x$, $\mathbb{P}[D(z_0) \geq D(z_1) \geq ... \geq D(z_K)|X = x] = 1$.

**Theorem 1.** (Th.1, B25 pp.17) Suppose that Assumption 4 is satisfied, then $\beta_{\text{tsls}}$ is WC if and only if the IV specification has RC.

Prop. 4 and Theorem 1 show that RC is both sufficient and necessary for the TSLS estimand to be weakly causal, where the necessity is novelly proven by B22 and B25, and is thus the main result of the paper. Without RC, the estimand includes baseline levels. Without monotonicity, some of the weights can be negative.

However, a weakly causal estimand does not necessarily have a clear counterfactual interpretation and may not be policy-relevant. In the following, we inspect how $\beta_{\text{rich}}$, the TSLS estimand under RC, relates to and deviates from the classical LATE framework.

We first consider a generalization of LATE under a binary instrument $Z \in \{0, 1\}$, which is from Angrist and Imbens (1995).

**Definition 5. Average causal response (ACR):**

$$\beta_{\text{acr}} := \mathbb{E}\left[Y(D(1)) - Y(D(0))|D(1) > D(0)\right] \tag{3}$$

If $D \in \{0, 1\}$, then $\beta_{\text{acr}}$ is LATE. B25 and Słoczyński (2024) pointed out that $\beta_{\text{acr}}$ may be the quantity that practitioners try to estimate. Taking covariates into consideration, we let $\beta_{\text{acr}}(x) := \mathbb{E}\left[Y(D(1)) - Y(D(0))|D(1) > D(0), X = x\right]$ denote the conditional ACR.

**Proposition 5.** (B25, pp.21-22) Given RC,

$$\beta_{\text{tsls}} = \frac{\mathbb{E}[Y(Z - \mathbb{E}[Z|X])]}{\mathbb{E}[Y(D - \mathbb{E}[Z|X])]} = \frac{\mathbb{E}[\text{Cov}(Y, Z|X)]}{\mathbb{E}[\text{Cov}(D, Z|X)]} =: \beta_{\text{rich}}. \tag{4}$$

$$\text{If } Z \in \{0, 1\}, \text{ then } \beta_{\text{acr}} = \mathbb{E}\left[\beta_{\text{acr}}(X)\frac{\mathbb{P}[D(1) > D(0)|X]}{\mathbb{P}[D(1) > D(0)]}\right], \text{ and} \tag{5}$$

$$\beta_{\text{rich}} = \mathbb{E}\left[\beta_{\text{acr}}(X)\frac{\mathbb{P}[D(1) > D(0)|X]\text{Var}(Z|X)}{\mathbb{E}\left[\mathbb{P}[D(1) > D(0)|X]\text{Var}(Z|X)\right]}\right] \tag{6}$$

Since B25 skipped formalizing Prop. 5, the proof is provided in Appendix A for clarity.

Intuitively, $\beta_{\text{rich}}$ is different from $\beta_{\text{acr}}$ in that its weights are influenced by $\text{Var}(Z|X)$. Partitioning the population by realizations of $X$, the subgroup with more variation in $Z$ is preferred by the estimand, which, when not accounted for, can produce misleading interpretations in practice. Note that (5) and (6) hold only under the binary instrument case. The exact interpretation of $\beta_{\text{rich}}$ remains an open question in general.

To conclude, for a TSLS estimand to be weakly causal—i.e. to provide an aggregate effect with the same sign as the local effects—it is **necessary** for the model to include covariates flexibly so that it satisfies the "rich covariates" condition. Combined with an assumption of treatment effect homogeneity or strong monotonicity, this guarantees weak causality. However, even under rich covariates, the estimand deviates from the traditional LATE framework, as its weights depend on $\mathrm{Var}(Z|X)$. These findings suggest that empirical researchers may need to adjust their modeling approaches and interpretation of the estimated effects accordingly.

## 3.3 Practice

B25 has provided a detailed guidance for empirical researchers to mitigate the rich covariates problem.

Firstly, the number of covariates included in the TSLS should be minimized. If theoretically justifiable, covariates should be avoided. Estimating $\beta_{\text{tsls}}$ without covariates produces the LATE of compliers. In contrast, the more covariates included, the more likely it is that RC is not satisfied.

If covariates are included, it is necessary to check if RC is satisfied. If a single instrument is included, this is equivalent to checking whether the linear regression $Z \sim X$ is correctly specified in terms of its functional form, which can be tested using the Ramsey (1969) RESET-test. It is implemented in `R` as `lmtest::resettest()` (Zeileis and Hothorn, 2002). The null hypothesis is equivalent to assuming that the specification has RC. Although not directly stated in B25, it is reasonable to also report the test at `type="regressor"` in addition to the default test on fitted values. The regressor test tests whether adding polynomials and interactions of some of the regressors significantly improves the model fit. In contrast, the default is a more general test and may not agree with the regressor non-linearity.

If the RESET-test rejects, then the TSLS estimand is, according to Theorem 1, not weakly causal. Since the bias cannot be quantified, it is recommended to also report the estimates from alternative methods. For example, the partial linear IV model (PLIV) in DDML works on the framework:

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

where $f(\cdot)$ and $g(\cdot)$ are estimated non-parametrically using base learners such as random forest and XGBoost. Since ML-based methods usually capture highly non-linear relationships among the regressors, $\beta_{\text{pliv}} = \beta_{\text{rich}}$. DDML-PLIV produces the estimand under rich covariates even when saturation is not possible. Additionally, if $D$ and $Z$ are binary, DDML-LATE estimates $\beta_{\text{acr}}$ and is thus LATE. However, the algorithm is empirically unstable and may sometimes degenerate. Both frameworks are implemented in the `ddml` package (Ahrens et al., 2024). We discuss some subtleties of DDML in Section 4.

Finally, practitioners should be cautious when interpreting the TSLS estimand. Even when the specification has rich covariates, the resulting $\beta_{\text{rich}}$ may be challenging to interpret, and reffering to it as a weighted average of LATEs is usually an oversimplification. Thus, B25 also suggests exploring other, semi- or non-parametric methods in IV analysis.

# 4 Application

In this section, I present practical examples related to the main result of B25. First, I run a simulation to illustrate the bias that arises when the specification does not satisfy RC. Using the simulated data, I then develop a framework for empirically approximating the weights in Prop. 1, which can be extended to real datasets. Next, I apply this framework to the adapted Card (1993) data. Finally, I examine an empirical paper by MacDonald et al. (2019), which is not covered in B25.

## 4.1 Simulation

To better illustrate the problem seen in Prop. 1 and 2, I begin by constructing a fixed data-generating process (DGP). The primary goals of this simulation are twofold. First, it provides a controlled setting where a direct comparison between $\beta_{\text{true late}}$, $\beta_{\text{tsls}}$, $\beta_{\text{saturated}}$, and $\beta_{\text{pliv}}$ is possible, so that the extent of the potential bias caused by the lack of RC is more visible. Second, while elements in the weights, such as $\mathbb{E}[\tilde{Z}|X]$ and $\mathbb{P}(g|x)$, are known once the DGP is known, they must be estimated in real data. Thus, the simulation offers a test-bed for developing an empirical approximation to the theoretical weights.

For simplicity, I omit the exact coefficients used in generating the data. The code for replication can be found in Appendix B.

**Definition 6.** Let $X = (X_1, X_2, ...)$ be a set of covariates.

1. int$(X)$ denotes the additive inclusion of all columns in X *plus* their interactions (pairwise, three-way, etc.) up to any desired order. All terms are included additively and no polynomials are involved.

2. $\text{poly}_d(X)$ denotes the additive inclusion of each column in $X$ in polynomial terms up to degree $d$, but without any cross-interactions among different columns.

**Example 1. (Simulation Procedure)**

**Step 1: Covariates.** Draw four i.i.d. "observed" normal random variables $n$ times

$$C_1,\ C_2,\ C_3,\ C_4\ \sim\ \mathcal{N}(0,1).$$

Bin each into quartiles (e.g. Q1 $<$ Q2 $<$ Q3 $<$ Q4) to obtain ordered categorical $X_1, X_2, X_3, X_4 \in \{\text{Q1}, \text{Q2}, \text{Q3}, \text{Q4}\}$. Let $X := (X_1, X_2, X_3, X_4)$. Since $X$ only contains categorical variables, the SW specification can be applied.

Additionally, there exists an "unobserved" variable $W\ \sim\ \mathcal{N}(\mu, \sigma)$.

**Step 2: Instrument *Z*.** Define some logistic index and sample $Z$ from a Bernoulli distribution, i.e. for the $i$-th individual,

$$\zeta(X_i)\ :=\ \text{int}(X_i),\ \ Z_i \sim \text{Ber}\big(\text{inv.logit}(\zeta(X_i))\big).$$

To replicate the B25 problem, it is necessary not to let RC be satisfied. This is checked through regressing $Z$ on $X$ and then performing the RESET-test. The null hypothesis is rejected.

**Step 3: Potential and Observed Treatment.** Define some logistic index

$$\delta_i(z) := \text{int}(X_i) \times W_i + \text{poly}_3(W_i), \quad z \in \{0, 1\}.$$

To rule out defiers, we set $\delta_i(1) \leftarrow \max\{\delta_i(0),\ \delta_i(1)\}$, and draw $D_i(z) \sim \text{Ber}(\text{inv.logit}(\delta_i(z)))$.

Finally set the observed treatment $D_i = Z_i\, D_i(1) + \big(1 - Z_i\big) D_i(0)$.

At this step, we know the potential treatment $D_i(z)$ of each individual. Therefore, we can label the **type** of each of them as AT, NT or CP accordingly.

At this and the following step, $W$ is included to make sure that $D$ is an endogeneous variable, so as to justify the use of IV.

The exogeneity of $Z$ in Assumption 2 can be checked by the regression $Z \sim X + \text{type}$, where type does not have significant effects. This check is only theoretical and cannot be applied in practice.

**Step 4: Potential and Observed Outcome.**

$$Y_i(d) := \text{int}(X_i)+\text{poly}_3(X_{1,i}+X_{2,i}+X_{3,i}+X_{4,i})+\text{poly}_3(W_i),\ \ Y_i = D_i\,Y_i(1)+\big(1-D_i\big)Y_i(0).$$

Then we get the true LATE as $\beta_{\text{true late}} = \mathbb{E}[Y_i(1) - Y_i(0) \mid \text{type}_i = \text{CP}]$.

**Step 5: Downstream Analysis.** With the full sample $\big\{(X_i,\ Z_i,\ D_i,\ Y_i)\big\}_{i=1}^{n}$ in hand, we compare the coefficient on $D$ from estimators: OLS, TSLS without covariates, TSLS with covariates, saturated TSLS, SW TSLS, DDML-PLIV, and DDML-LATE—a workflow also seen in Section 6 of B25. An exemplary result based on the set of coefficients used in the code is shown in Table 2.

| Method | Coefficient | Std. Error | Signif. | Other Info |
|---|---:|---:|:---:|---:|
| **True LATE** | **9.95** | – | – | – |
| OLS | 16.05 | 0.37 | *** | adj. $R^2 = 0.63$ |
| TSLS w/o X | -5.36 | 1.56 | *** | Weak inst. p=0, Wu-H p=0 |
| TSLS w. X | 4.51 | 1.10 | *** | Weak inst. p=0, Wu-H p=0 |
| Saturated TSLS | 9.66 | 1.10 | *** | Weak inst. p=0, Wu-H p=0 |
| TSLS, SW | 13.07 | 0.79 | *** | Weak inst. p=0, Wu-H p=0 |
| DDML-PLIV | 9.59 | 0.96 | *** | 100-fold, nnls |
| DDML-LATE | 9.66 | 1.32 | *** | 100-fold, nnls |

Table 2: Comparison of estimation results based on the simulated data. *Note:* '***' indicates $p < 0.001$. "Weak inst. p" and "Wu-H p" refer to the weak-instrument and Wu–Hausman test p-values, respectively. As suggested by B25, DDML uses XGBoost and Random Forest as base learners and is trained with 100 sample folds. The reported estimand is from the non-negative least squares (nnls) ensemble.

The significant Wu–Hausman tests confirm that endogeneity is present, explaining why both OLS and TSLS without covariates deviate substantially from the true LATE. In particular, OLS clearly overestimates the effect, whereas TSLS with covariates underestimates it—suggesting that negative weighting may be at play. By contrast, the saturated TSLS approximates the true LATE rather closely. The assignment of $Z$ is random given $X$, so that $\text{Var}(Z|X)$ is uniform across the dataset and the weights in Prop. 5 are almost negligible. An imbalanced assignment design can induce more discrepancy in $\beta_{\text{rich}}$. By contrast, the saturate-and-weight approach clearly induces many instruments bias, overestimating the effect just as OLS does. Finally, the DDML estimators track the LATE comparatively well, although they exhibit higher uncertainty and require tuning for better performance. During the experiment, the DDML-LATE algorithm was sensitive to CV-folds, sample folds, and base learners chosen. It is recommended to carefully tune the LATE estimation. In practical settings where saturation is infeasible, DDML provides a valuable correction over a basic "TSLS + covariates" setup.

**Step 6: Weights.** Recall Prop. 1, all components in the weights can be calculated as long as the DGP is known.

- $\mathbb{E}[\tilde{Z}D]^{-1}$ can be calculated from the observed data. Similarly, $\mathbb{L}[Z|X]$ comes from the fitted values of the linear regression $Z \sim X$.
- $\mathbb{E}[Z|X] = \text{inv.logit}(\zeta(X))$ due to the fact that each realization is drawn from the Bernoulli distribution. $\mathbb{E}[\tilde{Z}|X] = \mathbb{E}[Z|X] - \mathbb{L}[Z|X]$.
- $\mathbb{P}(G = g|X_i) = \begin{cases} \delta_i(0), & g = \text{AT} \\ \delta_i(1) - \delta_i(0), & g = \text{CP} \\ 1 - \delta_i(1), & g = \text{NT} \end{cases}$, as $\delta_i(z) = \mathbb{P}(D = 1|X_i, Z = z)$

Figure 2 shows the calculated weights. Clearly, compliers receive non-negative weights by design. The size of the weights are dependent on the specific covariate combination. Since $\mathbb{E}[\tilde{Z}|X]$ is usually not zero, AT and NT also receive weights, and for some covariate groups, AT and NT receive negative weights that are notable in size. This validates the findings of B25. Additionally, the calculated weights among CP have a Spearman correlation coefficient $r^{sp}_{\text{CP}} = 0.30,\ p = 10^{-6}$, while AT and NT are not significantly correlated with the conditional variance. This provides supplementary evidence for Prop. 5.

However, $\mathbb{E}[Z|X]$ and $\mathbb{P}(G = g|X_i)$ are inaccessible in real data and have to be estimated. $\mathbb{E}[Z|X]$ is estimated as the probability output of a nonparametric model that classifies $Z$ using $X$. Recall

$$\mathbb{P}(G = g|X_i) = \begin{cases} \mathbb{P}(D = 1|X_i, Z = 0), & g = \text{AT} \\ \mathbb{P}(D = 1|X_i, Z = 1) - \mathbb{P}(D = 1|X_i, Z = 0), & g = \text{CP} \\ 1 - \mathbb{P}(D = 1|X_i, Z = 1), & g = \text{NT} \end{cases},$$

$\mathbb{P}(G = g|X_i)$ can therefore be regarded as the probability output of a nonparametric model that classifies $D$ using $X$ and $Z = z$. We select the model class based on 10-fold cross validation (CV), where I compare the accuracy of GLM, Random Forest, and XGBoost, with hyperparameter tuning. Notably, XGBoost has more stable and accurate performance than a logistic regression, even when the targets $Z$ and $D$ are generated through a Bernoulli distribution. To avoid overfitting, I use the out-of-fold (OOF) predictions of the best model as its final output. In practice, there is no reason to expect

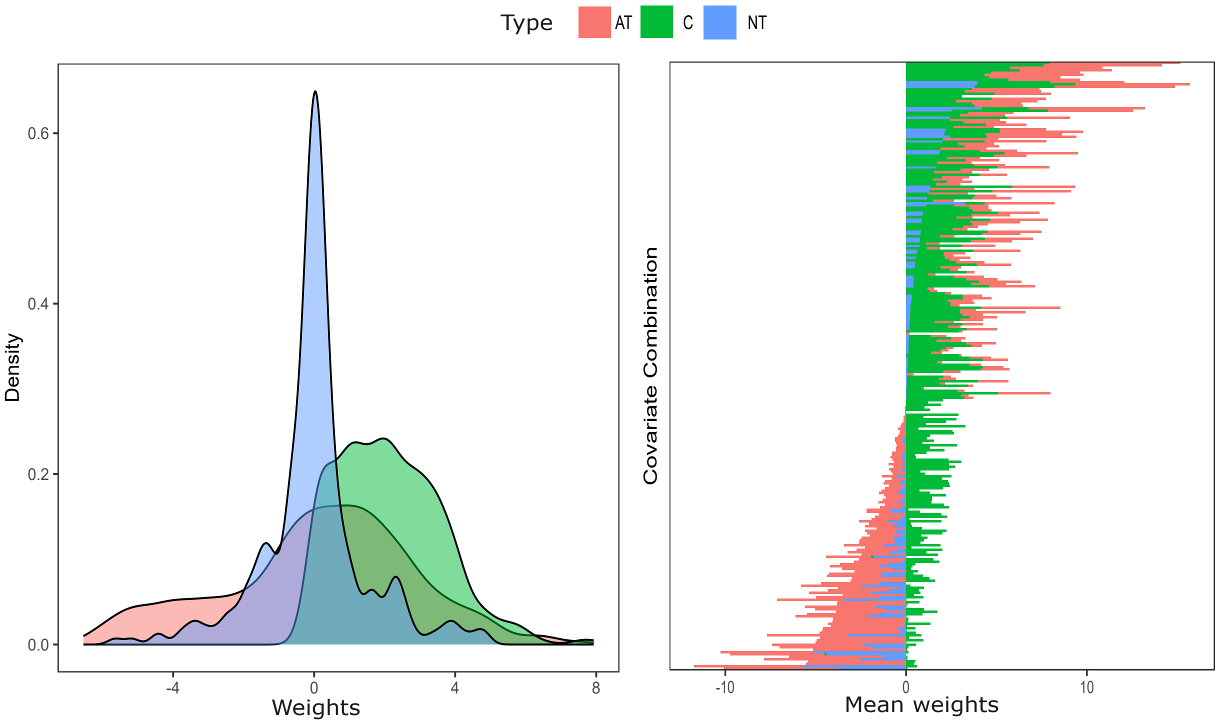

Figure 2: Weights calculated from the simulated data. *Note:* On the left is a density plot of weights colored by subpopulation type. On the right: $y$-axis is the different category combinations of $X$, such as $X = (Q_1, Q_2, Q_1, Q_1)$, the exact combinations are irrelavant and thus omitted. $x$-axis is the mean weights of the covariate combination, colored by type.

which model always outperforms. A CV-based model comparison should be conducted. For more robustness, repeated CV is recommended.

Figure 3 shows the estimated weights based on the simulated data. In this scenario, the models produce estimates that are similar to true weights in distribution, and according to the scatterplot, the mean deviation of the estimated weights from the true weights is not problematic.

In practice, the performance of this algorithm, especially whether it can correctly predict the sign of the weights, heavily depends on how well $Z$ is explained by $X$, which is not among the standard assumptions of IV analysis. The estimation can turn out noisy, if unobserved factors heavily influence the assignment of the instrument. Similarly, since the PLIV framework includes $Z = g(X) + v$, $\mathbb{E}[v|X] = 0$, the performance of DDML also depends on how well $X$ explains $Z$. For further discussion, see Section 5. Finally, we note that this algorithm is only applicable when the instrument and treatment have a **binary** support. For non-binary $D$ and $Z$, the estimation can be clumsy due to the inclusion of too many choice groups. For continuous $D$ or $Z$, the current definition of choice group

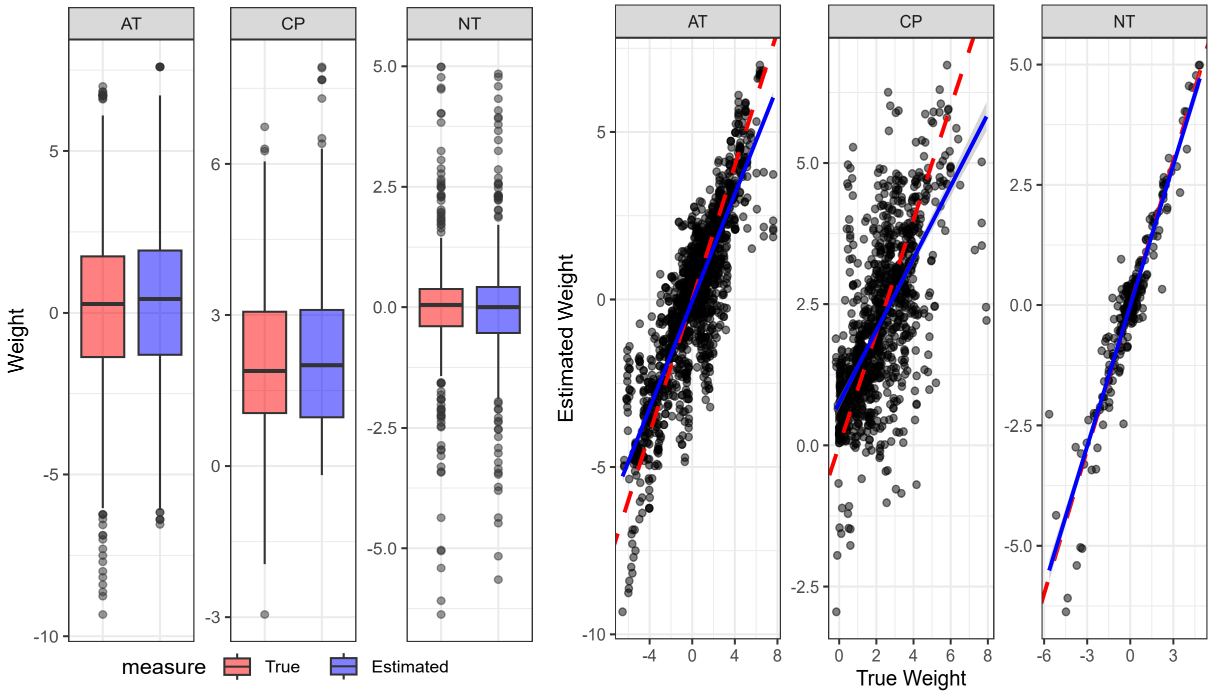

Figure 3: Comparison of estimated and true weights. *Note:* Facetted by subpopulation type. On the left is a boxplot of weights colored by true/estimated weights. On the right is a scatterplot. The red dotted line is $x = y$, and the blue line is the linear fit of the points.

must be defined differently. In those cases, although estimating $\mathbb{E}[\tilde{Z}|X]$ is feasible, it is not sufficient for investigating the effect of weights. According to Prop. 2, without RC, the estimand is level dependent, but the bias—the product of weights and baseline levels, is intricate to investigate.

## 4.2 Empirical Studies

In the following, I apply the Step 6 from Example 1 on empirical studies. First, I work on the adapted data from Card (1993). The original study investigates how education affects income among 3010 24–34-year-old men. In the dataset, $Y$ is log hourly wage, $D$ is years of education, and the binary $Z$ indicates whether a college is present in the local labor market. Covariates include years of professional experience, a race indicator for Black, indicators for living in the south of the US and for in urban areas, as well as variables regarding the respondents' past and family information. For simplicity, I applied a binary treatment, defined as an indicator of receiving more than 13 years of education, which is the median of the data, i.e. $D^* := \mathbb{I}(D > 13)$. $X$ includes four binary indicators: professional experience longer than 8 years (median), race indicator, living in the south, and in urban areas. The result of IV analysis is listed in Column (1) of Table 3. The discrepancy between OLS and IV estimands and the significant Wu-Hausman test justify the IV analysis. As the original paper argued, controlling for covariates is necessary. The RESET-test overwhelmingly rejects the null hypothesis, which is equivalent to the TSLS estimand not being weakly causal. However, $\beta_{\text{tsls}}$ is only minimally different from $\beta_{\text{saturated}}$ and $\beta_{\text{pliv}}$, indicating that in this specific case, negative weighting may not impose a noticeable bias. The SW estimand is clearly biased towards OLS. The DDML-LATE estimand differs from $\beta_{\text{pliv}}$ substantially, indicating that the variance-based weights as described in Prop. 5 should be accounted for when interpreting the causal effect of education, as the weights are, contentwise, irrelevant to the research question.

Next, I estimate the weights on $\beta_{\text{tsls}}$ as suggested above with 10-fold CV. In estimating $\mathbb{E}[Z|X]$, the logistic regression taking full interactions in $X$ had the best mean AUC at 0.75. When estimating $\mathbb{P}(G|X)$, random forest performed best with mean AUC at 0.76. The estimated weights are visualized in Figure 4. Notably, covariate groups with a high observation count are not necessarily assigned with large weights. Groups with "black" and "south" are more often than not negatively weighted. Although these weights do not substantially affect the value of the estimand, the resulting estimand is not weakly causal.

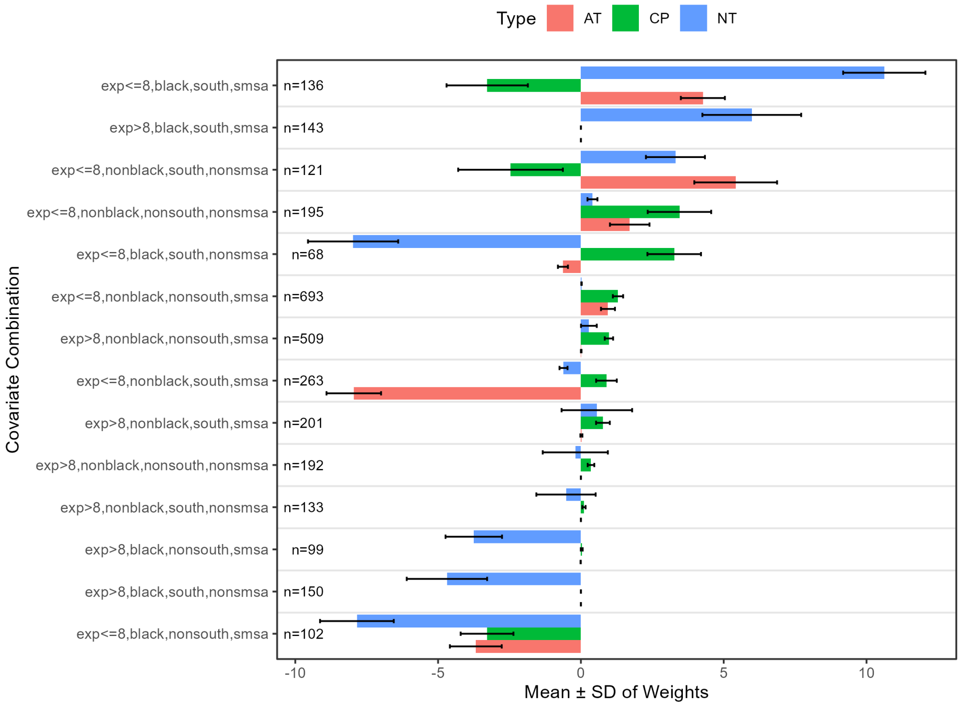

Figure 4: Estimated weights from Card (1993), Mean-SD plot. *Note:* the group "exp>8,black,nonsouth,nonsmsa" is not shown due to lack of observations (n=5).

Subsequently, I apply the above procedure on MacDonald et al. (2019), an empirical paper that is not investigated in B25. The paper studies whether an increased human-forest interaction leads to more Lyme disease incidence. The data is collected from 514 counties[1] in the US. The outcome $Y$ is a log Lyme disease incidence measure; $D$ is the percentage of population living near forest; $Z$ is a score measuring the local land use regulation. Covariates include an indicator for the year in which the survey is conducted, and three continuous variables: forest segmentation measure, mean patch area, and forest coverage. The county index is not included in the model but is used to calculate cluster-robust standard errors. Since variables are mostly continuous, estimating LATE and saturating in $X$ are not possible.

The result of IV analysis is listed in Column (2) of Table 3. Notably, the RESET-test on fitted values does not reject the null hypothesis, which indicates a lack of overall non-linearity in the regression of $Z$ on $X$. Although this does not sufficiently imply rich covariates, it indicates that the bias described in B25 may not be substantial in this scenario. This is reflected by the similar estimands yielded by TSLS with covariates and DDML-PLIV, which approximates $\beta_{\text{rich}}$. Note that the coefficients from TSLS with and without covariates are almost identical. It does not imply that including coefficients is not necessary. In fact, the first stage regression sees a substantial improvement in adjusted $R^2$ after including $X$ (from 0.18 to 0.38), which indicates that $X$ is contributing to the analysis. Considering the significant RESET-test on regressors, I experimented on including polynomials of $X$. The coefficient on $D$ remains consistent around 0.04 across different formulae. Overall, we can confidently say that the paper's main result is consistent and not severely affected by the rich covariates bias.

Although the rich covariates bias was not markedly problematic in both of the above cases, B25 has conducted a more extensive survey. I refer to Table 3 and Figure 4 in B25 for further details. In 10 out of 14 studies surveyed, RESET-test was significant, resulting in not weakly causal estimands. The bias, however, varied a lot across studies. In extreme cases, such as Nunn and Wantchekon (2011), the DDML-PLIV estimand was 75% less in magnitude than the original result, indicating a severe overestimation due to misspecification. In other cases including Card (1993), the bias was subtle (less than 10%).

[1] Only included complete cases.

|  | (1) Card (1993), adapted |  |  | (2) MacDonald et al. (2018) |  |  |
|---|---|---|---|---|---|---|
| **Measure** | **Coeff. on D** | **Std. Error** | **Signif.** | **Coeff. on D** | **Std. Error** | **Signif.** |
| OLS | 0.18 | 0.02 | *** | 0.008 | 0.004 | . |
| TSLS w/o X | 1.32 | 0.23 | *** | 0.037 | 0.009 | *** |
| TSLS w. X | 0.90 | 0.44 | * | 0.037 | 0.014 | ** |
| Saturated TSLS | 0.91 | 0.44 | * | – | – | – |
| TSLS, SW | 0.41 | 0.18 | * | – | – | – |
| DDML-PLIV | 0.93 | 0.44 | * | 0.035 | 0.010 | *** |
| DDML-LATE | 0.66 | 0.33 | * | – | – | – |
| **Diagnostics** | N: 3010 |  |  | N: 514 |  |  |
|  | RESET (fitted) p-val: 0.000 |  |  | RESET (fitted) p-val: 0.579 |  |  |
|  | RESET (regressor) p-val: NA |  |  | RESET (regressor) p-val: 0.017 |  |  |
|  | Weak Inst. p-val: 0.009 |  |  | Weak Inst. p-val: 0.000 |  |  |
|  | Wu–Hausman p-val: 0.033 |  |  | Wu–Hausman p-val: 0.000 |  |  |

Table 3: Estimates from Card (1993) and MacDonald et al. (2018). *Note:* Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. RESET-test on regressor is not applied to column (1) since all covariates are binary. Column (2) only includes complete cases. Std. errors of OLS and TSLS in column (2) are clustered by county index, whereas DDML SE are reported *as is*. The DDML specification is the same as in Table 2.

# 5 Subsequent Research

In this section, I examine how B22 and B25 have influenced subsequent theoretical and empirical works on TSLS. First, I replicated the search strategy described on page 10 (Section 2, B25), restricting publications to the period from January 2022 through February 2025 in the top five journals. This yielded 14 empirical studies utilizing TSLS, yet none cited B22 or B25. Consequently, I broadened the search to include all papers citing B22, without limiting to specific journals. Via Google Scholar, I identified 11 widely cited empirical articles, two closely related methodological papers, and two review articles. For the empirical articles, I focus on the following aspects:

1. whether and how they addressed rich covariates;
2. whether and how they addressed the monotonicity assumption;
3. how they interpreted the resulting estimand;
4. whether additional IV methods are reported in addition to TSLS.

Table 4 presents a consise summary of the 11 papers. A more detailed table is provided in Appendix B.

Goldfarb et al. (2022) is a well-cited review on marketing research with quasi-experiments. When comparing IV methods such as TSLS, the authors briefly mentioned B22, referring to its critiques as *"subtleties"* when interpreting the estimand. This serves as a representative example of the general attitude among practitioners. Specifically, none of the 11 papers reported a RESET-test to validate rich covariates as recommended. While most of the 11 papers did not include additional IV methods suggested by B22 or B25, such as DDML, Conzelmann et al. (2023) also reported estimands based on SW and JIVE.

Regarding rich covariates, 4 out of 11 papers approached the specification in the traditional manner: they performed robustness checks by adding interaction or polynomial terms of covariates, arguing that their main result was consistent if different specifications yielded similar estimands, or at least estimands with a same sign (Clemens et al., 2022; Heldring et al., 2022; Hener, 2022; Schwab and Singh, 2024). While exploring alternative specifications helps validate the consistency of the analysis, a non-exhaustive survey does not work in the same way as an evidence-based RESET-test. In contrast, Almarzooq et al. (2023) explicitly included rich covariates as a model assumption. Another paper by Chyn et al. (2022) avoided including covariates after discussing B22. Gallen et al. (2023) fully accounted for covariate richness by first matching observations based on covariates to create homogeneous subgroups, and then specifying a fully saturated model. Meanwhile, the remaining 4 papers (Agan et al., 2023; Conzelmann et al., 2023; Garin et al., 2024; Gray-Lobe et al., 2022) addressed covariates by implementing what the methodological review by Borusyak et al. (2024) called "formula instruments". In these studies, the instrument is constructed via a known function of certain observed covariates and exogenous shocks, which are assumed randomly assigned conditional on these covariates. Typically, categorical variables are used in the formula. Therefore, saturating in these variables is possible and is equivalent to having rich covariates, while other covariates can be included linearly. However, this approach demands strong theoretical justification and is not commonly viable. Conceptually, formula instruments align with what we in Section 4.1 mentioned—$X$ should explain $Z$ well to reduce bias in IV analysis.

The 11 papers handle monotonicity in a similar way. In studies with binary $D$ and $Z$, the authors either implicitly assume no-defiers, or justify the assumption in theory. In those working on non-binary variables, strong conditional monotonicity is assumed as in Assumption 4. An exception is Agan et al. (2023), in which the authors argued that strong monotonicity was too restrictive and cannot be tested. Instead, they followed Frandsen et al. (2023)[2] and tested "average monotonicity", i.e. $\mathrm{Cov}(D(z), Z) \geq 0$, which is weaker than Assumption 4. According to Theorem 1, without strong monotonicity, the TSLS estimand may involve negative weighted individual treatment effects.

When interpreting the TSLS estimand, 5 studies cautiously termed it as "weighted, subgroup-specific treatment effects" (Almarzooq et al., 2023; Conzelmann et al., 2023; Gallen et al., 2023; Garin et al., 2024; Hener, 2022). Additionally, Conzelmann et al. (2023) argued that the weights are "potentially not policy-relevant". Among the 4 papers that referred to the estimand as (weighted average of) complier LATEs, Chyn et al. (2022) and Gray-Lobe et al. (2022) followed the assumptions in B25, while the specifications in Agan et al. (2023) and Heldring et al. (2022) partly deviated from Theorem 1. The remaining two papers did not specify their interpretation.

Finally, B22 have inspired further methodological research. In Alvarez et al. (2024), the authors proposed a framework of a dynamic regression discontinuity design, adopting the definition of "weak causality" in B22 to clarify the exact causal interpretation of their method. As an extension to B25, which focuses on variables with a discrete support, Alvarez and Toneto (2024) derived weights of the TSLS estimand given a continuous instrument and a binary treatment, using rich covariates and strong monotonicity as assumptions.

[2] The paper is published in AER. For lack of accessibility, I could not examine its content in detail. However, it seems to have relevant methodological contribution and should be considered in future survey.

| Author | Year | Rich covariates | Monotonicity | LATE interpretation | Additional methods |
|---|---|---|---|---|---|
| Agan et al. | 2023 | formula instrument | Average monotonicity | Yes | SW, JIVE |
| Almarzooq et al. | 2023 | RC as an assumption | No-defiers | No | None |
| Chyn et al. | 2022 | avoided covariates | No-defiers | Yes | None |
| Clemens et al. | 2022 | robustness checks | No-defiers | NA | None |
| Conzelmann et al. | 2023 | formula instrument | No-defiers | No | None |
| Gallen et al. | 2023 | matching and saturation | No-defiers | No | None |
| Garin et al. | 2024 | formula instrument | Strong conditional monotonicity | No | None (discussed JIVE) |
| Gray-Lobe et al. | 2023 | formula instrument | Strong conditional monotonicity | Yes | None |
| Heldring et al. | 2023 | robustness checks | No-defiers | Yes | MTE |
| Hener | 2022 | robustness checks | Strong conditional monotonicity | No | None |
| Schwab & Singh | 2024 | robustness checks | No-defiers | NA | None |

Table 4: Summary of Subsequent Research. *Note:* in column LATE interpretation: “Yes” stands for an interpretation as a weighted average of complier LATEs. “No” as weighted, subgroup-specific treatment effects. “NA” means not specified.

# 6 Discussion

Overall, B25 focuses on the sufficient and necessary conditions for addressing the misspecification bias of TSLS with covariates. It contributes to the previous discussion by Evdokimov and Kolesár (2018) and Słoczyński (2024) primarily by establishing the necessity of rich covariates. In contrast to existing works, which typically assume rich covariates, B25 demonstrates through empirical examples that this condition is often violated. The "weakly causal" definition, unique to B25, can inspire future research that examines the causal interpretation of an estimand. Furthermore, B25 provides practical guidance for practitioners, including reporting RESET-test and DDML estimands.

Nevertheless, B25 has several limitations. First, the estimand decomposition in Prop. 2 is restricted to $D$ and $Z$ with finite support. For continuous treatment and instrument, although Theorem 1 still holds, it remains unclear how the weights are defined.

Second, even when we are aware of a theoretical estimand decomposition and thus the presence of misspecification bias, its contribution to empirical studies may be limited. If $D$ and $Z$ are non-binary, the weights can be challenging to estimate, and even then it is difficult to quantify the misspecification bias. As displayed in Section 4, the bias has a considerable variance among empirical studies, ranging from subtle to overwhelming. In Figure 4, the negative weights have large absolute values, yet exert little influence on the estimand. Future studies could focus on exploring the theoretical or empirical conditions that produce substantial misspecification bias.

Third, the weakly causal estimand $\beta_{\text{rich}}$ lacks a clear causal interpretation. As discussed in Section 3.2, it can currently be viewed as a positively weighted average of covariate-specific treatment effects. Prop. 5 shows it is dependent on Var(Z|X) under binary treatment and instrument, and is not a naive aggregate of complier LATEs, contrary to some practitioners' expectations. However, for more general cases, its interpretation requires further investigation.

Finally, B25 is limited to a single instrument, as the derivation of main propositions rely on Equation 1. Future work could extend the analysis of misspecification bias to scenarios with multiple instruments.

In this paper, I have reviewed earlier and subsequent research centered around B25, synthesized its main theoretical contributions, and enhanced those insights through both simulation and empirical applications. Nonetheless, some notable limitations remain. First, although the simulation in Example 1 demonstrates how misspecification bias can arise, future work could expand it to multiple functional forms or sample scenarios, thereby testing the robustness of its result. Second, the current weight estimation strategy is restricted to binary $Z$ and $D$. Extending it to general scenarios can benefit empirical research. It is also worth exploring how the interplay of weights and individual treatment effects finally affect the magnitude of misspecification bias. Lastly, the survey on subsequent work in Section 5 focuses on those citing B22 or B25. It would be valuable to explore whether other recent TSLS-based studies are adopting novel practices to mitigate misspecification bias.

In sum, the presence of covariates in TSLS is not benign. While some subsequent empirical studies regard the problem as "subtle", researchers are increasingly aware of it. New methods, including formula instruments, have been developed to mitigate misspecification bias by ensuring covariate richness. Moreover, novel non-parametric estimation strategies, such as DDML, can complement traditional TSLS. Although the classical LATE framework is being re-examined, continuing research will guide empirical studies toward identifying the causal effect of interest more rigorously.

# A Appendix

## A.1 Proof of Prop. 5

**Proof.** First, rewrite Equation (1) using the definition of $\tilde{Z}$:

$$\beta_{\text{tsls}} = \frac{\mathbb{E}[Y(Z - \mathbb{L}[Z|X])]}{\mathbb{E}[Y(D - \mathbb{L}[Z|X])]} = \frac{\mathbb{E}[Y(Z - \mathbb{E}[Z|X])]}{\mathbb{E}[Y(D - \mathbb{E}[Z|X])]}$$

due to RC. Consider the numerator, using law of iterated expectations, it is

$$\mathbb{E}\left[\mathbb{E}[Y(Z - \mathbb{E}[Z|X])|X]\right] = \mathbb{E}[\mathbb{E}[YZ|X] - \mathbb{E}[Y|X]\mathbb{E}[Z|X]] = \mathbb{E}[\text{Cov}(Y, Z|X)]$$

according to the definition of covariance. Doing the same on the denominator yields (4). Next, rewrite (3) into

$$\beta_{\text{acr}} = \frac{\mathbb{E}\left[Y(D(1)) - Y(D(0))\mathbb{I}[D(1) > D(0)]\right]}{\mathbb{P}[D(1) > D(0)]}$$

Rewrite the numerator with law of iterated expectations:

$$\mathbb{E}\left[Y(D(1)) - Y(D(0))\mathbb{I}[D(1) > D(0)]\right] = \mathbb{E}\left[\mathbb{E}[Y(D(1)) - Y(D(0))\mathbb{I}[D(1) > D(0)]|X]\right]$$

Note that

$$\begin{aligned}\mathbb{E}[Y(D(1)) - Y(D(0))\mathbb{I}[D(1) > D(0)]|X] =& \mathbb{E}[Y(D(1)) - Y(D(0))|D(1) > D(0), X] \\ & \cdot \mathbb{P}[D(1) > D(0)|X] \\ =& \beta_{\text{acr}}(X)\mathbb{P}[D(1) > D(0)|X]\end{aligned}$$

Dividing by $\mathbb{P}[D(1) > D(0)]$ yields (5).

Finally, rewrite (4). Applying Equation (21) in Proof of Proposition 2, Appendix A pp.35, B25[3], under monotonicity, we have

$$\begin{aligned}\text{Cov(D, Z|X)} =& \mathbb{P}(G = \text{CP}|X) \cdot \mathbb{E}[Z|X](1 - \mathbb{E}[Z|X]) \\ =& \mathbb{P}[D(1) > D(0)|X]\text{Var}(Z|X) \\ \text{Cov(Y, Z|X)} =& \text{CATE}(\text{CP}, X) \cdot \text{Cov(D, Z|X)} \\ =& \mathbb{E}[Y(1) - Y(0)|D(1) > D(0), X]\text{Cov(D, Z|X)}\end{aligned}$$

[3] Also seen in Imbens and Angrist (1994).

$$=\beta_{\text{acr}}(X)\mathbb{P}[D(1) > D(0)|X]\text{Var}(Z|X)$$

Taking expectations on both terms, we have

$$\begin{aligned}\beta_{\text{rich}} =& \frac{\mathbb{E}[\beta_{\text{acr}}(X)\mathbb{P}[D(1) > D(0)|X]\text{Var}(Z|X)]}{\mathbb{E}[\mathbb{P}[D(1) > D(0)|X]\text{Var}(Z|X)]} \\ =& \mathbb{E}\left[\beta_{\text{acr}}(X)\frac{\mathbb{P}[D(1) > D(0)|X]\text{Var}(Z|X)}{\mathbb{E}\left[\mathbb{P}[D(1) > D(0)|X]\text{Var}(Z|X)\right]}\right]\end{aligned}$$

□

# B Electronic appendix

The code and figures are available in the GitHub Repository: `https://github.com/Yc-Han/Causal_TSLS`.

**List of Contents:**

- a detailed table for subsequent research papers: `subsequent_research.xlsx`;
- `R` codes for replication in `\R`, inclduing `card`, `lyme`, and `simulation`;
- original and additional plots in `\output`.

**Data Availability:**

- The Card (1993) data is downloaded from the Mixtape repository on GitHub: `https://github.com/scunning1975/mixtape`.
- The MacDonald et al. (2019) data is available on DRYAD: `https://doi.org/10.5061/dryad.p7t9289`.

# References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models, *Journal of Econometrics* **113**(2): 231–263.

Agan, A., Doleac, J. L. and Harvey, A. (2023). Misdemeanor prosecution*, *The Quarterly Journal of Economics* **138**(3): 1453–1505.

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

Almarzooq, Z. I., Song, Y., Dahabreh, I. J., Kochar, A., Ferro, E. G., Secemsky, E. A., Major, J. M., Farb, A., Wu, C., Zuckerman, B. and Yeh, R. W. (2023). Comparative effectiveness of percutaneous microaxial left ventricular assist device vs intra-aortic balloon pump or no mechanical circulatory support in patients with cardiogenic shock, *JAMA Cardiology* **8**(8): 744–754.

Alvarez, L. A. and Toneto, R. (2024). The interpretation of 2sls with a continuous instrument: A weighted late representation, *Economics Letters* **237**: 111658.

Alvarez, L., Orestes, V. and Silva, T. (2024). Corporate effects of monetary policy: Evidence from central bank liquidity lines, *Technical report*, MIT Working Paper.

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity, *Journal of the American statistical Association* **90**(430): 431–442.

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

Athey, S., Tibshirani, J. and Wager, S. (2019). Generalized random forests, *The Annals of Statistics* **47**(2): 1148 – 1178.

Blandhol, C., Bonney, J., Mogstad, M. and Torgovitsky, A. (2022). When is tsls actually late?, *Technical report*, National Bureau of Economic Research Cambridge, MA.

Blandhol, C., Bonney, J., Mogstad, M. and Torgovitsky, A. (2025). When is tsls actually late?, *Technical report*, National Bureau of Economic Research Cambridge, MA.

Borusyak, K., Hull, P. and Jaravel, X. (2024). Design-based identification with formula instruments: a review, *The Econometrics Journal* **28**(1): 83–108.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling, *Working Paper 4483*, National Bureau of Economic Research.

Chamberlain, G. and Imbens, G. (2004). Random effects estimators with many instrumental variables, *Econometrica* **72**(1): 295–306.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters, *The Econometrics Journal* **21**(1): C1–C68.

Chyn, E., Haggag, K. and Stuart, B. A. (2022). The effects of racial segregation on intergenerational mobility: Evidence from historical railroad placement, *Working Paper 30563*, National Bureau of Economic Research.

Clemens, J., Hoxie, P. G. and Veuger, S. (2022). Was pandemic fiscal relief effective fiscal stimulus? evidence from aid to state and local governments, *Working Paper 30168*, National Bureau of Economic Research.

Conzelmann, J. G., Hemelt, S. W., Hershbein, B., Martin, S. M., Simon, A. and Stange, K. M. (2023). Skills, majors, and jobs: Does higher education respond?, *Working Paper 31572*, National Bureau of Economic Research.

Evdokimov, K. S. and Kolesár, M. (2018). Inference in instrumental variables analysis with heterogeneous treatment effects, *Technical report*, Princeton University.

Frandsen, B., Lefgren, L. and Leslie, E. (2023). Judging judge fixed effects, *American Economic Review* **113**(1): 253–277.

Gallen, Y., Joensen, J. S., Johansen, E. R. and Veramendi, G. F. (2023). The labor market returns to delaying pregnancy, *Available at SSRN 4554407* .

Garin, A., Koustas, D. K., McPherson, C., Norris, S., Pecenco, M., Rose, E. K., Shem-Tov, Y. and Weaver, J. (2024). The impact of incarceration on employment, earnings, and tax filing, *Working Paper 32747*, National Bureau of Economic Research.

Goldfarb, A., Tucker, C. and Wang, Y. (2022). Conducting research in marketing with quasi-experiments, *Journal of Marketing* **86**(3): 1–20.
**URL:** *https://doi.org/10.1177/00222429221082977*

Gray-Lobe, G., Pathak, P. A. and Walters, C. R. (2022). The long-term effects of universal preschool in boston*, *The Quarterly Journal of Economics* **138**(1): 363–411.

Heldring, L., Robinson, J. A. and Vollmer, S. (2022). The economic effects of the english parliamentary enclosures, *Working Paper 29772*, National Bureau of Economic Research.

Hener, T. (2022). Noise pollution and violent crime, *Journal of Public Economics* **215**: 104748.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects, *Econometrica* **62**(2): 467–475.
**URL:** *http://www.jstor.org/stable/2951620*

MacDonald, A. J., Larsen, A. E. and 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. and 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.

Schwab, S. D. and Singh, M. (2024). How power shapes behavior: Evidence from physicians, *Science* **384**(6697): 802–808.

Słoczyński, T. (2024). When should we (not) interpret linear iv estimands as late?

Zeileis, A. and Hothorn, T. (2002). Diagnostic checking in regression relationships, *R News* **2**(3): 7–10.
**URL:** *https://CRAN.R-project.org/doc/Rnews/*