# Design-Based Inference for Two-State Least Squares Estimation

Robin Danko

October 2025

**Abstract**

In the typical approach to inference in the social sciences, researchers assume that a negligibly small sample is drawn from a large population. While natural in many applications, it is less natural in other instances such as when doing inference on statewide data when data on all 50 states is available; in this case the sample does not differ from the population. In this article, I apply a design-based approach to analyze the population variance of the 2SLS estimator. Design-based uncertainty explicitly considers the unknown counterfactual outcomes under alternative treatment schemes. I derive standard errors of the 2SLS estimator that consider both design-based and sampling-based uncertainty. I show that these new standard errors are generally smaller than the usual infinite population sampling-based standard errors and provide conditions under which they coincide.

## 1 Introduction

When considering sampling schemes, a typical approach to the social sciences assumes that a sample comprises only a small proportion of the population of interest. Indeed, it is assumed that this proportion is so small that it is approximate to sampling from an infinite population. Under this frame, uncertainty is only generated by random sampling from a population. This is natural in many settings: if a researcher is studying individual level data from the one percent public-use U.S. census sample, only a small proportion of the population of interest is observed, so the infinite-population model is a good approximation. In many settings, however, a significant fraction of the population is observed, this assumption is not as natural. For instance, Manski and Pepper [5] study the effect of right-to-carry laws on crime rates. Their data contains information on the laws and homicide rights of every U.S. state. In other words, this data samples every unit in the population. In this case, uncertainty is generated by the counterfactual outcomes, not the sampling scheme; the researcher cannot know both the realized outcome of a state having a right-to-carry law and not having a right-to-carry law at any one time.

This paper shows that when a significant proportion of the population of interest is observed in a sample, the usual approach to heteroskedasticity-robust inference on instrumental variable estimates in the social sciences uses standard errors that are overly conservative because these standard errors only consider sampling-based uncertainty and ignore design-based uncertainty: uncertainty generated by unknown counterfactuals for sampled units. Abadie et. al (2020) [1] demonstrated this result for heteroskedasticity-robust standard errors used for inference on OLS estimators. These authors also demonstrate a similar result for inference on a binary treatment when treatment assignment is clustered (2023) [2].

In the spirit of design-based uncertainty, this paper explicitly models the counterfactual outcomes generated by the treatment and the instrument. This follows the literature on randomized experiments (Rosenbaum 2002) [6], but this work considers these counterfactuals in regression analyses and observational studies as well. All other variables are treated as fixed characteristics of the population. This is justified by counterfactual interpretations that focus on one concrete population: the question of interest is how varying the treatment levels changes the outcomes for that population with the observed set of characteristics, not the outcomes of a hypothetical population with a completely different set of characteristics.

This paper builds on the growing literature demonstrating the importance of considering design-based uncertainty and presents an alternative variance estimate for IV estimates that is less conservative than the standard approach only considering sampling-based uncertainty. In many situations in which a significant proportion of the population is observed, the entire population is still not observed. Thus, this framework also integrates sampling-based uncertainty where it is appropriate.

## 2 A Simple Example

Suppose the population of interest has size $n$. $N$ observations are sampled from this population. The variable $R_i \in \{0,1\}$ records if a unit was sampled ($R_i = 1$) or not ($R_i = 0$).

For each sampled unit, the researcher observes an outcome $Y_i$, a *causal* regressor $W_i$, a binary instrument $Z_i$, and an *attribute* $X_i$. The instrument is assumed to be random and generates potential treatment, while the attribute is assumed to be fixed.

Suppose the instrument generates potential values of the causal variable $W_i^*(0)$ and $W_i^*(1)$. Considering the classical example of returns to education studied in Card 1993 [4], $W_i^*(1)$ and $W_i^*(0)$ could indicate college graduation with or without living near a 4-year university respectively. The realization of the causal variable is

$$W_i = W_i^*(Z_i) = \begin{cases} W_i^*(1) \text{ if } Z_i = 1 \\ W_i^*(0) \text{ if } Z_i = 0 \end{cases}$$

The instrument is assumed to only influence $Y_i^*$ through $W_i$, that is $Y_i^*(W_i, Z_i) = Y_i^*(W_i)$. Therefore, there are two potential outcomes for each unit $Y_i^*(1)$ and $Y_i^*(0)$. Moreover, randomness in both the potential treatment and the potential outcome only comes from the randomness of the instrument. Returning to the returns to education example, $Y_i^*(1)$ and $Y_i^*(0)$ could be earnings with or without a college degree. The realized outcome is

$$Y_i = Y_i^*(W_i) = \begin{cases} Y_i^*(1) \text{ if } W_i = 1 \\ Y_i^*(0) \text{ if } W_i = 0 \end{cases}$$

Define $\mathbf{Y}, \mathbf{Y}^*(1), \mathbf{Y}^*(0), \mathbf{R}$, $\mathbf{Z}$, $\mathbf{W}, \mathbf{W}^*(1)$ and $\mathbf{W}^*(0)$ to be the population-length vectors with $i$th element equal to $Y_i$, $Y_i^*, Y^*(0), R_i, Z_i, W_i, W^*(1)$, and $W^*(0)$ respectively. The researcher observes $Y_i, Z_i$, and $W_i$ for every sampled unit and observe $R_i$ for every unit.

Note that the above definition of the potential outcomes requires some assumptions; for instance, it embeds the "no interference" or Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1978) [7].

**Assumption 2.1** (SUTVA)**.**

1. *If* $W_i = W_i^{'}$ *then* $\mathbf{Y}_i^*(\mathbf{W}) = \mathbf{Y}_i^*(\mathbf{W}^{'})$
2. *If* $Z_i = Z_i^{'}$ *then* $\mathbf{W}_i^*(\mathbf{Z}) = \mathbf{W}_i^*(\mathbf{Z}^{'})$ .

That is, both the potential treatment and potential outcome functions for an observation only depend on the realized treatment and instrument value for that observation and are independent of the treatment and instrument values for the other units in the population. This is an unnatural assumption in some contexts, such as vaccine treatments where there tend to be spillover effects.

The attribute does not generate potential values, which is to say that it does not create counterfactual outcomes and the researcher is not interested in causal inference of its impact on the outcome. The potential outcomes themselves are non-stochastic attributes. This is similar to a control variable in a regression about which the researcher is agnostic as to whether or not the variable is endogenous. Although counterfactual values of these variables cannot be considered, they assist in analysis on causal variables as will be shown in greater detail in Section 3. In the current section I drop the attributes for simplicity.

Continuing to follow the language of Abadie et al. (2018), I call estimands *causal* if they are functions of the full set of values $(\mathbf{Y}^*(1), \mathbf{Y}^*(0), \mathbf{Z}, \mathbf{R}, \mathbf{W}^*(1), \mathbf{W}^*(1))$ for every unit in the population, whether or not they are sampled. This full set of values also includes all levels of the potential outcomes which are not observed even for sampled units. In particular, a causal estimand of interest is the following:

$$\theta^{\text{causal}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(Y^*\left(W_i^*(1)\right) - Y^*\left(W_i^*(0)\right)\right)}{\frac{1}{n}\sum_{i=1}^{n} W_i^*(1) - W_i^*(0)}.$$

The above estimand is causal because it is a function of both levels of the potential outcomes, not just what is observed in a sample. Just because an estimand is "causal" by this definition doesn't mean that it has a meaningful ceteris paribus interpretation that is useful for practitioners. Under the following assumptions, the above estimand is interpreted as the average treatment effect on the compliers, that is, those units for whom $W_i^*(1) \neq W_i^*(0)$

**Assumption 2.2** (Monotonicity)**.** *Either*

$$W_i^*(1) \geq W_i^*(0)\ \forall i \ \ or\ \ W_i^*(1) \leq W_i^*(0)\ \forall i$$

*must be true.*

This assumption is discussed in Angrist and Imbens (2015) [3] as an identifying assumption for causal effects. It supposes that the instrument can only influence the treatment in one direction. For instance, in the returns to schooling example, people who are born within 25 miles of a four-year university must receive at least as much schooling as they would have completed had they lived farther away from a four-year university.

By contrast, if an estimand can be written as a function of only the realized outcomes $(\mathbf{Y}, \mathbf{W}, \mathbf{Z})$, without dependence on either the sampling or the potential outcomes, it is called a *descriptive* estimand. That is, if the researcher observed the realized outcomes for every unit of the population, a descriptive estimand would be known with certainty. Let $n_z$ denote the number of units in the population such that $Z_i = z$. Then a descriptive estimand of interest is:

$$\theta^{\text{desc}} = \frac{\frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)Y_i}{\frac{1}{n_1}\sum_{i=1}^{n} Z_i W_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)W_i}.$$

That is, $\theta^{\text{desc}}$ is simply the Wald estimator computed using the entire population, not just the sampled units. If the entire population was observed, this estimand would be known with certainty. Let $N_z$ denote the number of units in the sample such that $Z_i = z$. The estimator of interest is the same Wald estimator computed using only the sampled units:

$$\hat{\theta} = \frac{\left(\frac{1}{N_1}\sum_{i=1}^{n} R_i Z_i Y_i - \frac{1}{N_0}\sum_{i=1}^{n} R_i(1-Z_i)Y_i\right)}{\left(\frac{1}{N_1}\sum_{i=1}^{n} R_i Z_i W_i - \frac{1}{N_0}\sum_{i=1}^{n} R_i(1-Z_i)W_i\right)}$$

Comparing this estimator to $\theta^{\text{causal}}$ and the full set of potential outcomes, two types of data are missing. For every unsampled unit, the researcher has no information about any of the variables. This generates what Abadie et al. (2018) call *sample-based uncertainty*. For every sampled unit the unrealized potential outcomes are not observed, labeled *design-based uncertainty*.

**Assumption 2.3** (Random Sampling without replacement)**.**

$$Pr(\mathbf{R} = \mathbf{r}) = 1\Big/\binom{n}{N}$$

*for all $n$-vectors* $\mathbf{r}$ *with* $\sum_{i=1}^{n} r_i = N$

**Assumption 2.4** (Random Assignment)**.**

$$Pr(Z_i = 1|\mathbf{R}) = \frac{1}{n_1}$$

*for all $n$-vectors* $\mathbf{z}$ *with* $\sum_{i=1}^{n} Z_i = n_1$

**Assumption 2.5** (Instrument Relevance)**.**

$$\frac{1}{n}\sum_{i=1}^{n}\left(W_i^*(1) - W_i^*(0)\right) \neq 0$$

Calculating the variance of this estimator is not straightforward since it is the ratio of two random variables. However, asymptotic analysis admits tools like the delta method to approximate this variance in large samples. So far, however, I have only considered estimands that are tied to a population with fixed size $n$. Therefore, to perform asymptotics, I consider a sequence of populations indexed by their size $n$. Then, $Y_{n,i}$ would indicate the observed outcome for unit $i$ in population $n$. Similarly, define $Z_{n,i}, W^*_{n,i}$, and $R_{n,i}$. From this sequence of populations, consider a sequence of estimands $\theta_n^{\text{causal}}$ and $\theta_n^{\text{desc}}$ and a sequence of estimators $\hat{\theta}_n$. Then, I can establish the consistency of this new estimator by arguing that $(\hat{\theta}_n - \theta_n^{\text{causal}}) \xrightarrow{p} 0$ and make similar asymptotic arguments. Modeling a sequence of increasing populations allows both the sample and the population to grow together (possibly at different rates) while preserving a notion of sampling a non-negligible fraction of units from a population.

# 3 The General Case

I focus on a setting with a scalar outcome, a collection of regressors, and a set of instruments. A regressor is either a causal regressor or an attribute. Attributes are defined to be fixed characteristics of a population and therefore do not admit counterfactual analysis. Causal regressors are defined as random variables and admit counterfactuals. For instance, consider a vaccine trial conducted on a group of participants aged 55-65 where treatment is determined randomly. The ages of the participants are attributes determined prior to the assignment mechanism; treatment assignment is causal.

## 3.1 Setup

Consider a sequence of populations indexed by finite population size $n$. Each unit, indexed $i$, has a set of fixed attributes $X_{n,i}$, a fixed potential outcome function $Y^*_{n,i}(\cdot)$, and a $1 \times K$ fixed potential treatment function $W^*_{n,i}(\cdot)$. The potential outcome function generates realized outcomes determined by the causal variables $W_{n,i}$. These realized outcomes are denoted $Y_{n,i} = Y^*_{n,i}(W_{n,i})$. The potential treatment function generates realized levels of the causal variables determined by the $1 \times L$ set of instruments $Z_{n.i}$. These realized outcomes are denoted $W_{n,i} = W^*_{n,i}(Z_{n,i})$. $W_{n,i}$ and $Z_{n,i}$ are real-valued column vectors, and $Y_{n,i}$ is scalar. All of these variables can be continuous, discrete, or mixed.

Each population has an associated sample with sampling vector $\mathbf{R}$. $R_{n,i} = 1(R_{n,i} = 0)$ indicates that unit $i$ of population $n$ is sampled (not sampled). For each sampled unit, the researcher observes $(Y_{n,i}, W_{n,i}, Z_{n,i}, X_{n,i})$.

Define for each population the matrices $\mathbf{Y}_n$, $\mathbf{W}_n$, $\mathbf{Z}_n$, $\mathbf{R}_n$,$\mathbf{X}_n$ , and $\mathbf{Y}^*_n$ to be matrices that contain their respective variables for every unit in the population.

**Assumption 3.1** (Assignment Mechanism)**.** *The assignment of the instrument $Z_{n,1}, ..., Z_{n,n}$ are jointly independent and independent of the sampling assignment $R_{n,1}, ..., R_{n,n}$. but not necessarily identically dis-*

*tributed (i.n.i.d.).*

Throughout this paper I will work with a transformation of the instruments denoted $V_{n,i}$. Assume that $\sum_{i=1}^{n} X_{n,i}X_{n,i}'$ is full-rank. Then,

$$V_{n,i} = Z_{n,i} - \Xi_n X_{n,i}, \text{ where } \Xi_n = \left(\sum_{i=1}^{n} E[Z_{n,i}]X_{n,i}'\right)\left(\sum_{i=1}^{n} X_{n,i}X_{n,i}'\right)^{-1}.$$

$V_{n,i}$ is the IV version of partialling out the linear relationship with the controls from the instruments in a Frisch-Waugh-Lovell sense. Since $\Xi_{n,i}X_{n,i}$ are nonstochastic and $Z_{n,i}$ are i.n.i.d., $V_{n,1},...,V_{n,n}$ are i.n.i.d. as well.

**Assumption 3.2** (Random sampling)**.** *i) There exists a sequence of sampling probabilities $\rho_n$ such that*

$$Pr(\boldsymbol{R}_n = r) = \rho_n^{\sum_{i=1}^{n} r_i}(1-\rho_n)^{n-\sum_{i=1}^{n} r_i}$$

*for all vectors $r$ of length $n$, where $r_i \in \{0,1\}$. ii) The sequence of sampling rates $\rho_n$ satisfies $n\rho_n \to \infty$ and $\rho_n \to \rho \in [0,1]$*

In addition to a usual random sampling condition, the above assumption requires that the expected sample size increases as the population size increases and that the expected sample size converges to some proportion of the population $\rho$. As a special case, $\rho$ is allowed to be 0, in which case the usual robust IV standard errors are robust.

I impose a regularity assumption on the moments of the variables.

**Assumption 3.3** (Moment conditions)**.** *There exists some $\delta > 0$ such that that the sequences*

$$\frac{1}{n}\sum_{n=1}^{n} E[|Y_{n,i}|^{4+\delta}], \quad \frac{1}{n}\sum_{n=1}^{n} E[||V_{n,i}||^{4+\delta}], \quad \frac{1}{n}\sum_{i=1}^{n} ||X_{n,i}||^{4+\delta}, \quad \frac{1}{n}\sum_{i=1}^{n} E[||W_{n,i}||^{4+\delta}]$$

*are uniformly bounded.*

Define

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix}\begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix}', \qquad \Omega_n = \frac{1}{n}\sum_{i=1}^{n} E\left[\begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix}\begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix}'\right]$$

.

$\Omega_n = E[P_n]$, with expectation taken over the joint distribution of $W_n$ and $V_n$. I also define versions of $W_n$ and $\Omega_n$ incorporating sampling:

$$\tilde{P}_n = \frac{1}{N}\sum_{i=1}^{n} R_{n,i}\begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix}\begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix}', \qquad \tilde{\Omega}_n = \frac{1}{N}\sum_{i=1}^{n} R_{n,i} E\left[\begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix}\begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix}'\right].$$

where $\tilde{\Omega}_n = E[\tilde{P}_n | R_n]$. It is convenient to partition these matrices to refer to sections of them later:

$$P_n = \begin{pmatrix} P_n^{YY} & P_n^{YW} & P_n^{YV} & P_n^{YX} \\ P_n^{WY} & P_n^{WW} & P_n^{WV} & P_n^{WX} \\ P_n^{VY} & P_n^{VW} & P_n^{VV} & P_n^{VX} \\ P_n^{XY} & P_n^{XW} & P_n^{XV} & P_n^{XX} \end{pmatrix}$$

I will refer to analogous partitions for $\Omega_n$, $\tilde{P}_n$, and $\tilde{\Omega}_n$.

**Lemma 3.1.** *Suppose Assumptions 3.1, 3.2, and 3.3 hold. Then* $\lim_{n\to\infty} \tilde{P}_n - \Omega_n \xrightarrow{p} 0$

**Assumption 3.4** (Convergence of Moments). $\lim_{n\to\infty} \Omega_n \to \Omega$. $\Omega$ *is full rank*

**Assumption 3.5** (Instrument relevance).

1. $E\left[\sum_{i=1}^{n} \frac{1}{n} V_{n,i} V'_{n,i}\right]$ *has full rank*
2. $E\left[\sum_{i=1}^{n} \frac{1}{n} W_{n,i} V'_{n,i}\right]$ *has rank* $K$.

Note that the assumption of instument relevance is aggregated across units; the instrument need not affect every unit in the same was even in expectation. Some units are allowed to be never or alwaystakers. This assumption only requires that the treatment status of some units is affected by exposure to the instrument.

Define

$$\Pi_n = \Omega_n^{WV}(\Omega_n^{VV})^{-1}, \quad \tilde{\Pi}_n = \tilde{\Omega}_n^{WV}(\tilde{\Omega}_n^{VV})^{-1}$$

$$Q_n = P_n^{WV}(P_n^{VV})^{-1}, \quad \tilde{\Pi}_n, \quad \tilde{Q}_n = \tilde{P}_n^{WV}(\tilde{P}_n^{VV})^{-1}.$$

$Q_n$ and $\tilde{Q}_n$ are the coefficients obtained from the first-stage regression of the treatment on the instruments using the entire population and the sample respectively. $\Pi_n$ and $\tilde{\Pi}_n$ are the moment versions of these coefficients. As a consequence of Lemma 3.1 $\operatorname{plim}_{n\longrightarrow\infty} \tilde{Q}_n - \Pi_n = 0$

Using the same methodology as the previous section, define two estimands of interest, which exist with probability approaching 1 under the conditions imposed above.

$$\begin{pmatrix} \theta_n^{desc} \\ \gamma_n^{desc} \end{pmatrix} = \begin{pmatrix} P_n^{VW}(P_n^{VV})^{-1}P_n^{WV} & P_n^{XV}Q_n \\ Q_nP_n^{VX} & P_n^{XX} \end{pmatrix}^{-1} \begin{pmatrix} Q_nP_n^{VY} \\ P_n^{XY} \end{pmatrix}$$

$$\begin{pmatrix} \theta_n^{desc} \\ \gamma_n^{causal} \end{pmatrix} = \begin{pmatrix} \Omega_n^{VW}(\Omega_n^{VV})^{-1}\Omega_n^{WV} & \Omega_n^{XV}\Pi_n \\ \Pi_n\Omega_n^{VX} & \Omega_n^{XX} \end{pmatrix}^{-1} \begin{pmatrix} \Pi_n\Omega_n^{VY} \\ \Omega_n^{XY} \end{pmatrix}.$$

In particular, $\begin{pmatrix} \theta_n^{\text{desc}} \\ \gamma_n^{\text{desc}} \end{pmatrix}$ is the usual 2SLS estimator calculated using every unit in the population, not just the sample. Note that in the just-identified case when $K = L$

$$\begin{pmatrix} \theta_n^{desc} \\ \gamma_n^{desc} \end{pmatrix} = \begin{pmatrix} P_n^{VW}P_n^{VX} \\ P_n^{XW}P_n^{XX} \end{pmatrix}^{-1} \begin{pmatrix} P_n^{VY} \\ P_n^{XY} \end{pmatrix}$$

$$\begin{pmatrix} \theta_n^{causal} \\ \gamma_n^{causal} \end{pmatrix} = \begin{pmatrix} \Omega_n^{VW}\Omega_n^{VX} \\ \Omega_n^{XW}\Omega_n^{XX} \end{pmatrix}^{-1} \begin{pmatrix} \Omega_n^{VY} \\ \Omega_n^{XY} \end{pmatrix}$$

which is the usual IV estimator.

Define residuals calculated with respect to both estimands

$$\begin{aligned} e_{n,i}^{\text{causal}} &= Y_{n,i} - W_{n,i}'\theta_n^{\text{causal}} - X_{n,i}'\gamma_n^{\text{causal}}, \\ e_{n,i}^{\text{desc}} &= Y_{n,i} - W_{n,i}'\theta_n^{\text{desc}} - X_{n,i}'\gamma_n^{\text{desc}}. \end{aligned}$$

$\theta_n^{\text{desc}}$ and $\theta_n^{\text{causal}}$ correspond to the following orthogonality conditions respectively:

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} V_{n,i} \\ X_{n,i} \end{pmatrix} e_{n,i}^{\text{desc}} = 0, \quad \frac{1}{n}\sum_{i=1}^{n} E\left[\begin{pmatrix} V_{n,i} \\ X_{n,i} \end{pmatrix} e_{n,i}^{\text{causal}}\right] = 0.$$

In the just-identified case, the 2SLS estimator uniquely corresponds to this orthogonality condition. In the overidentified case when $L > K$, the choice of 2SLS is not unique. Other choices within a GMM framework are possible. This paper analyzes the properties of the 2SLS estimator $\hat{\theta}_n$ obtained by the following matrix representation:

$$\begin{pmatrix} \hat{\theta}_n \\ \hat{\gamma}_n \end{pmatrix} = \begin{pmatrix} \tilde{P}_n^{VW}(\tilde{P}_n^{VV})^{-1}\tilde{P}_n^{WV} & \tilde{Q}_n\tilde{P}_n^{VX} \\ \tilde{P}_n^{XV}\tilde{Q}_n & \tilde{P}_n^{XX} \end{pmatrix}^{-1} \begin{pmatrix} \tilde{Q}_n\tilde{P}_n^{VY} \\ \tilde{P}_n^{XY} \end{pmatrix}.$$

The $E[Z_{n,i}]$s are unknown, and so $V_{n,i}$ cannot be computed. However, the above estimator is asymptotically equivalent to the following estimator which is always feasible. Define

$$\hat{V}_{n,i} = Z_{n,i} - \Bigg(\bigg(\sum_{i=1}^{n} R_{n,i}V_{n,i}X_{n,i}\bigg)\bigg(\sum_{i=1}^{n} R_{n,i}(X_{n,i}X_{n,i}'\bigg)^{-1} R_{n,i}X_{n,i}\Bigg).$$

as the feasible estimator of $V_{n,i}$ and

$$\begin{aligned} \hat{Q}_{n,i} &= \frac{1}{n}\sum_{i=1}^{n} R_{n,i}\hat{V}_{n,i}'\left(\frac{1}{n}\sum_{i=1}^{n} R_{n,i}\hat{V}_{n,i}\hat{V}_{n,i}\right)^{-1}\hat{V}_{n,i} \\ \hat{W}_{n,i} &= \hat{Q}_{n,i}\hat{V}_{n,i}. \end{aligned}$$

Then I use an alternative estimator $(\tilde{\theta}_n, \tilde{\gamma}_n)$, defined as:

$$\begin{pmatrix} \tilde{\theta}_n \\ \tilde{\gamma}_n \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} R_{n,i}\hat{W}_{n,i}\hat{W}_{n,i}' & \frac{1}{n}\sum_{i=1}^{n} R_{n,i}\hat{W}_{n,i}X_{n,i}' \\ \frac{1}{n}\sum_{i=1}^{n} X_{n,i}\hat{W}_{n,i}' & \frac{1}{n}\sum_{i=1}^{n} R_{n,i}X_{n,i}X_{n,i}' \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} R_{n,i}\hat{W}_{n,i}Y_{n,i}' \\ \frac{1}{n}\sum_{i=1}^{n} R_{n,i}X_{n,i}Y_{n,i}' \end{pmatrix}.$$

These estimators satisfy $\lim_{n\longrightarrow\infty}(\hat{\theta}_n - \tilde{\theta}_n)$, and $\tilde{\theta}$ is always feasible to compute as long as full-rank conditions hold in the sample. In the just-identified case, when $K = L$ (when the number of causes and instruments are the same), the 2SLS estimator takes the form

$$\begin{pmatrix} \hat{\theta}_n \\ \hat{\gamma}_n \end{pmatrix} = \begin{pmatrix} \tilde{P}_n^{VW} & \tilde{P}_n^{VX} \\ \tilde{P}_n^{XW} & \tilde{P}_n^{XX} \end{pmatrix}^{-1} \begin{pmatrix} \tilde{P}_n^{VY} \\ \tilde{P}_n XY \end{pmatrix}.$$

## 3.2 Causal interpretation of estimands

By the definitions offered above, $\theta_n^{\text{causal}}$ is a causal estimand because it is a function of the population moments, but that does not necessarily imply that the estimand has an useful causal effect interpretation. In this section, I present conditions under which $\theta_n^{\text{causal}}$ is informative about the underlying potential outcome function.

**Assumption 3.6** (Expected Assignment for Instrument)**.** *There exists a sequence of functions $g_n$ such that*

$$E[Z_{n,i}] = g_n(X_{n,i})$$

*and a sequence of matrices $C_n$ such that for all $x$*

$$g_n(x) = C_n(x)$$

*for all $n$ large enough*

First, note that if the instrument is randomly assigned then the above condition is satisfied as long as $X_{n,i}$ contains an intercept since $E[Z_{n,i}]$ is a constant. Assumption 3.6 allows for the instrument to depend on the attributes, while restricting the mean-dependence of $E[Z_{n,i}]$ to be eventually linear in $X_{n,i}$.

**Assumption 3.7.**

$$Y_{n,i}^{*}(W_{n,i}) = W_{n,i}^{'}\theta_{n,i} + \xi_{n,i}$$

*where $\theta_{n,i}$ and $\xi_{n,i}$ are nonstochastic.*

The vector $\theta_{n,i}$ represents the causal effect of increasing the corresponding value of $W_{n,i}$ by one unit for unit $i$ in population $n$. The non-stochastic nature of $\xi_{n,i}$ is worth remarking on, in particular when considering under which conditions the consistency of OLS won't hold. Abadie et. al (2020) [1] require that the following condition holds, similar to 3.6.

**Assumption 3.8** (Assumption 7 in Abadie et. al)**.**

$$E[W_{n,i}] = h_n(X_{n,i})$$

*and a sequence of matrices* $B_n$ *such that for all* $x$

$$h_n(x) = B_n(x)$$

*for all* $n$ *large enough*

Suppose instead that $E[W_{n,i}] = h_n(X_{n,i}) + f(\xi_{n,i})$, that is, allow some dependence between the treatment assignment and the non-stochastic error term. In this case, the OLS estimator does not estimate a useful causal estimand and IV is required. The authors in provide a specific case in which $\theta^{causal}$ is not a weighted average of the treatment effects.

**Theorem 3.1.** *Suppose assumptions 3.1-3.6 hold. Then, for all* $n$ *large enough,*

$$\theta_n^{causal} = \left(\Omega_n^{VW}\left(\Omega_n^{VV}\right)^{-1}\Omega_n^{WV}\right)^{-1}\Omega_n^{VW}\left(\Omega_n^{VV}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}E[V_{n,i}W_{n,i}^{'}]\theta_{n,i}\right)$$

*In the just-identified case*

$$\theta_n^{causal} = [\Omega_n^{VW}]^{-1}\sum_{i=1}^{n}E[V_{n,i}W_{n,i}^{'}]\theta_{n,i}.$$

First, note that if $V_{n,i}$ and $W_{n,i}$ are identically distributed, $\theta^{\text{causal}} = \frac{1}{n}\sum_{i=1}^{n}\theta_{n,i}$. That is, it can be interpreted as a simple average causal effect. In the following section, I discuss more tractable interpretations of $\theta_n^{\text{causal}}$ allowing $V_{n,i}$ and $W_{n,i}$ to be i.n.i.d.

## 3.3 Interpretation of Estimands

As aruged in the first section, when $W_{n,i}$ and $Z_{n,i}$ are both binary, $\theta_n^{\text{causal}}$ can be interpreted as the average treatment effect for the compliers, ie. for those observations for whom the instrument is relevant.

These results can be extended based on the work by Angrist and Imbens 1995 [3]. Suppose now that the treatment takes a finite number of values $W_{n,i} = 0, 1, 2, ...J$. I also suppose that every level of the treatment has a well-defined, fixed potential outcome for each individual.

$$Y_{n,i} = Y^*_{n,i} = \begin{cases} Y^*_{n,i}(0) & \text{if } W_{n,i} = 0 \\ & \vdots \\ Y^*_{n,i}(J) & \text{if } W_{n,i} = J. \end{cases}$$

In particular, the returns to education literature falls within this setting. In this case, suppose that each observational unit has a potential wage $Y^*_{n,i}(j)$ that corresponds to what their wage would be if they had completed $j$ levels of education. Continue to assume that the instrument is dyadic. Each observational unit has treatment level corresponding to both potential values of the treatment, $W^*_{n,i}(Z_{n,i})$. As before, a monotonicity assumption must be imposed for interpretation.

**Assumption 3.9** (Monotonicity assumption for multi-valued treatment)**.** $W_i(1) \geq W_i(0)$ *or* $W_i(0) \geq W_i(1)\ \forall i$

Define the following values

$$\omega_j = \frac{\sum_{i=1}^n \mathbf{1}(W_i(1) \geq j > W_i(0))}{\sum_{j=1}^J \sum_{i=1}^n \mathbf{1}(W_i(1) \geq j > W_i(0))}$$

$$\beta_j = \frac{1}{n}\sum_{i=1}^n Y_{i,j} - Y_{i,j-1}$$

$$\beta = \sum_{j=1}^J \omega_j \beta_J$$

**Lemma 3.2.** *Under assumptions 2.1-2.6 and assumption 2.8, for large enough n*

$$\theta_n^{causal} = \beta$$

That is, $\theta_n^{\text{causal}}$ can be interpreted as the treatment effect again on the compliers. In this case, compliers are those units for whom the instrument is relevant to their treatment.

## 3.4 Interactions Among Endogenous and Exogenous Variables

In special cases, interaction variables can improve interpretation of estimands and they are easily implemented in this framework. Suppose for simplicity of notation there is one treatment of interest suspected to be endogenous, one instrument, and one attribute. Without loss of generality, assume the attribute $X_{n,i}$

has mean 0. Suppose that

$$Y^*_{n,i}(W_{n,i}) = W_{n,i}\theta_n + W_{n,i}X_{n,i}\delta_n + X_{n,i}\gamma_n + \xi_{n,i} = W_{n,i}\theta_{n,i} + \xi_{n,i}$$

where $\theta_{n,i} = \theta_n + X_{n,i}\delta$. Observe that this is a special case of assumption 3.7. Moreover,

$$\theta_n = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial Y^*_{n,i}}{\partial W}.$$

Therefore, $\theta_n$ can be interpreted as the average effect of the treatment across all units in the population. If estimation proceeds without incorporating the interactive effect, the estimand that can be estimated will be the weighted average discussed in Theorem 3.1.

Suppose that instead one wanted to use $Z_{n,i}X_{n,i}$ as an instrument for $W_{n,i}X_{n,i}$. By assumption 3.6,

$$E[Z_{n,i}X_{n,i}] = E[Z_{n,i}]X_{n,i} = h_n(X_{n,i})X_{n,i} = B_n X^2_{n,i}$$

where the last equality holds for all sufficiently large $n$. Therefore, by including $X^2_{n,i}$ as an attribute, $Z_{n,i}X_{n,i}$ satisfies assumption 3.6 and is a valid instrument. By using the vector of instruments $(Z_{n,i}, Z_{n,i}X_{n,i})$ as instruments for the vector of treatments $(W_{n,i}, W_{n,i}X_{n,i})$, $\theta^{causal} = \theta$, $\delta^{causal} = \delta$ and both of these effects can be estimated consistently. Now instead assume the potential outcome takes the form

$$Y^*_{n,i}(W_{n,i}) = W_{n,i}\theta_n + W_{n,i}X_{n,i}\delta_{n,i} + X_{n,i}\gamma_n + \xi_{n,i} = W_{n,i}\theta_{n,i} + \xi_{n,i}$$

where $\theta_{n,i} = \theta_n + X_{n,i}\delta_{n,i}$, $\theta_n$ is identified by using the same procedure described above.

The identified parameter for the interaction term $\delta^{causal}$ will be the weighted representation discussed in Theorem 3.1. In this case, although the entire relationship of the outcome with the treatment cannot be identified, the average treatment effect is identified.

## 3.5 The Asymptotic Distribution of the IV Estimator

In this section, I analyze the IV estimator as an estimator of the estimands $\theta_n^{\text{causal}}$ and $\theta_n^{\text{desc}}$.

Define residuals calculated using the population causal estimands:

$$\epsilon_{n,i} = Y_{n,i} - W_{n,i}'\theta_n^{\text{causal}} - X_{n,i}'\gamma_n^{\text{causal}}$$

Here, it is not required to consider distinct residuals calculated using $\theta_n^{\text{desc}}$ since in the limit: $\theta_n^{\text{desc}} - \theta_n^{\text{causal}} \xrightarrow{p} 0$.

Note that the usual orthogonality condition $E[V_{n,i}\epsilon_{n,i}]$ can vary across all $i$, and in particular may not always be 0. Indeed, if $E[V_{n,i}\epsilon_{n,i}] = 0$ for all $i$, the usual EHW standard errors are not conservative; the alternative estimator proposed in this paper is only an improvement if this is not the case. However, as argued above $\frac{1}{n}\sum_{i=1}^n E\left[\begin{pmatrix} V_{n,i} \\ X_{n,i} \end{pmatrix}\epsilon_{n,i}\right] = 0$. Define the following limits of the population variance,

$$\Delta^{\text{cond}} = \lim_{n\longrightarrow\infty} \frac{1}{n}\sum_{i=1}^n \text{var}(V_{n,i}\epsilon_{n,i})$$

and the following expectation

$$\Delta^{\text{ehw}} = \lim_{n\longrightarrow\infty} \frac{1}{n}\sum_{i=1}^n E[\epsilon_{n,i}^2 V_{n,i}V_{n,i}'].$$

By definition, $\Delta^{\text{ehw}} - \Delta^{\text{cond}}$ is the following limit

$$\Delta^{\mu} = \lim_{n\longrightarrow\infty} \frac{1}{n}\sum_{i=1}^n E[V_{n,i}\epsilon_{n,i}]E[V_{n,i}\epsilon_{n,i}]'$$

and is positive semidefinite.

**Assumption 3.10** (Existence of Limits)**.** *$\Delta^{cond}$ and $\Delta^{ehw}$ exist and are positive definite.*

**Theorem 3.2.** *Suppose Assumptions 2.1-2.6, 2.9, and 2.10 hold, let* $\Gamma = \lim_{n\longrightarrow\infty}\left(\Omega_n^{VW}\left(\Omega_n^{VV}\right)^{-1}\Omega_n^{WV}\right)^{-1}\Omega_n^{VW}\left(\Omega_n^{VV}\right)^{-1}$, *Then,*

$$\sqrt{N}(\hat{\theta}_n - \theta_n^{causal}) \xrightarrow{d} \mathcal{N}\left(0, \Gamma^{-1}\left(\rho\Delta^{cond} + (1-\rho)\Delta^{ehw}\right)\Gamma^{-1}\right)$$

$$\sqrt{N}(\hat{\theta}_n - \theta_n^{desc}) \xrightarrow{d} \mathcal{N}\left(0, \Gamma^{-1}\left((1-\rho)\Delta^{ehw}\right)\Gamma^{-1}\right)$$

An immediate implication of the above theorem is that when $\rho = 0$, when the sample size is a negligible fraction of the population of interest, the asymptotic variance reduces to the usual sandwich variance $\Gamma^{-1}\Delta^{\text{ehw}}\Gamma^{-1}$ for both $\theta_n^{\text{causal}}$ and $\theta_n^{\text{desc}}$. Therefore, the standard setting where the sample is assumed to be from a very large or infinite population is a special case of the above theorem. However, whenever $\rho > 0$, the difference between the $\Delta^{\text{ehw}}$ and the asymptotic population variance of $\sqrt{N}(\theta_n^{\text{causal}} - \hat{\theta})$ is $\rho\Gamma^{-1}\Delta^{\mu}\Gamma^{-1}$ which is positive semidefinite.

The case where the entire population is observed is covered by the case $\rho = 1$. In this case, if a descriptive perspective is taken, the standard error will be 0, which is sensible since $\hat{\theta}_n = \theta_{\text{desc}}$ for any $n$. Therefore, taking account of the causal nature of a research problem is important.

## 3.6 The Variance Under Correct Specification

**Assumption 3.11** (Constant Treatment Effect)**.**

$$Y_{n,i}^{*} = w^{'}\theta_n + \xi_{n,i}$$

In this case, $\epsilon_{n,i}^{\text{causal}} = \xi_{n,i} - X_{n,i}^{'}\gamma^{\text{causal}}$. Every term in this expression is non-stochastic. Therefore, $E[V_{n,i}\epsilon_{n,i}] = E[V_{n,i}]\epsilon_{n,i} = 0$ and $\Delta^{\mu} = 0$. This implies the following result.

**Theorem 3.3.**

$$\sqrt{N}(\hat{\theta}_n - \theta_n^{causal}) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1}\Delta^{ehw}\Gamma^{-1})$$

*for all* $\rho$.

Therefore, the usual EHW variance estimator is robust. Notice that theorem 3 holds only when every treatment has a constant effect and doesn't always apply when one treatment effect is constant and another is heterogeneous. In particular, when considering the model

$$Y_{n,i}^{*}(W_{n,i}) = W_{n,i}\theta_n + W_{n,i}X_{n,i}\delta_{n,i} + X_{n,i}\gamma + \xi_{n,i} = W_{n,i}\theta_{n,i} + \xi_{n,i}$$

with instrument vector $(Z_{n,i}, Z_{n,i}X_{n,i})$, neither element of $E[V_{n,i}^{'}\epsilon_{n,i}]$ is generally 0. Therefore, the EHW variance specification is still conservative.

# 4 Estimating the Variance

In this section, I discuss estimating the variance for the above estimands. For convenience, define $V^{\text{causal}} = \Gamma^{-1}(\rho\Delta^{\text{cond}} + (1-\rho)\Delta^{\text{ehw}})\Gamma^{-1}$, $V^{desc} = (1-\rho)\Gamma^{-1}\Delta^{ehw}\Gamma^{-1}$, and $V^{\text{ehw}} = \Gamma^{-1}\Delta^{\text{ehw}}\Gamma^{-1}$.

There are four components to these asymptotic variances, $\rho$, $\Gamma$, $\Delta^{\text{ehw}}$, and $\Delta^{\text{cond}}$. $\rho$ can be estimated as $\hat{\rho}_n = \frac{N}{n}$. To estimate $\Gamma$, estimate $\Xi_n$ and as

$$
\begin{aligned}
\hat{\Xi}_n &= \left(\sum_{i=1}^{n} R_{n,i} Z_{n,i} X_{n,i}'\right)\left(\sum_{i=1}^{n} R_{n,i} X_{n,i} X_{n,i}'\right) \\
\hat{\Lambda}_n &= \left(\sum_{i=1}^{n} R_{n,i} W_{n,i} X_{n,i}'\right)\left(\sum_{i=1}^{n} R_{n,i} X_{n,i} X_{n,i}'\right).
\end{aligned}
$$

Then estimate $\Gamma$ by using the following sample average:

$$
\hat{\Gamma}_n = \left(\left(\frac{1}{N}\sum_{i=1}^{n} R_{n,i}\hat{V}_{n,i}W_{n,i}'\right)\left(\frac{1}{N}\sum_{i=1}^{n} R_{n,i}\hat{V}_{n,i}\hat{V}_{n,i}'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{n} R_{n,i}W_{n,i}\hat{V}_{n,i}'\right)\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{n} R_{n,i}\hat{V}_{n,i}W_{n,i}'\right)\left(\frac{1}{N}\sum_{i=1}^{n} R_{n,i}\hat{V}_{n,i}\hat{V}_{n,i}'\right)^{-1}
$$

$\Delta^{\text{ehw}}$ can also be estimated by estimating the residuals for all the sampled units $\hat{\epsilon}_{n,i} = Y_{n,i} - W_{n,i}'\hat{\theta}_n - Z_{n,i}'\hat{\gamma}_n$. Then estimate $\Delta^{ehw}$ as

$$
\hat{\Delta}_n^{\text{ehw}} = \frac{1}{N}\sum_{i=1}^{n} R_{n,i}(Z_{n,i} - \hat{\Xi}_n X_{n,i})\hat{\epsilon}_{n,i}^2(Z_{n,i} - \hat{\Xi}_n X_{n,i})'. \tag{1}
$$

Then the large sample variance, $V^{\text{ehw}}$ can be estimated as

$$
\hat{V}_n^{\text{ehw}} = \hat{\Gamma}_n^{-1}\hat{\Delta}_n^{\text{ehw}}\hat{\Gamma}_n^{-1}
$$

**Lemma 4.1.** *Suppose assumption 2.1-2.6 and 2.9 hold with* $\delta = 4$. *Then,*

$$
\hat{V}_n^{ehw} \xrightarrow{p} V^{ehw}
$$

$V^{\text{causal}}$ is not straightforward to estimate because it includes $\Delta^{\text{cond}}$. This cannot be estimated because it relies on the expectations $E[V_{n,i}\epsilon_{n,i}]$, which cannot be estimated consistently because only one observation for each unit is possible. While each individual $E[\epsilon_{n,i}^2 V_{n,i} V_{n,i}']$ similarly cannot be estimated, the average across all individuals can be estimated by the formula provided in 1.

Often, researchers use a conservative estimator of $\Delta^{\text{cond}}$, $\Delta^{\text{ehw}}$ because all of the the potential outcomes for a single unit cannot be observed. However, this paper proposes an improvement upon this conservative estimator by using the attributes. The proposed estimator replaces the expectations $E[V_{n,i}\epsilon_{n,i}]$, which cannot be consistently estimated, with predictors from a least squares projection of estimates of $V_{n,i}\epsilon_{n,i}$ on the attributes $X_{n,i}$. Let $\hat{V}_{n,i} = Z_{n,i} - \hat{\Xi}X_{n,i}$ and define

$$\hat{G}_n = \left(\frac{1}{N}\sum_{i=1}^{n} R_{n,i}\hat{V}_{n,i}\hat{\epsilon}_{n,i}X_{n,i}'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{n} R_{n,i}X_{n,i}X_{n,i}'\right).$$

**Assumption 4.1.**

$$\frac{1}{n}\sum_{i=1}^{n} E[V_{n,i}\epsilon_{n,i}]X_{n,i}'$$

*has a limit.*

The above assumption ensures that $\hat{G}_n$ converges. Then, I propose the following estimator similar to the one proposed in [1]:

$$\hat{\Delta}_n^z = \frac{1}{N}\sum_{i=1}^{n} R_{n,i}(\hat{V}_{n,i}\hat{\epsilon}_{n,i} - \hat{G}_n X_{n,i})(\hat{V}_{n,i}\hat{\epsilon}_{n,i} - \hat{G}_n X_{n,i})'.$$

This estimator uses $\hat{G}_n X_{n,i}$ instead of a consistent estimator of $E[V_{n,i}\epsilon_{n,i}]$. The proposed estimator is still conservative. However, in the next lemma I show that this approach can explain some of the variance in $\hat{V}_{n,i}\hat{\epsilon}_{n,i}$, and so the estimator $\hat{\Delta}_n^Z$ is smaller than $\hat{\Delta}_n^{\text{ehw}}$ in a matrix sense.

**Lemma 4.2.** *Suppose assumption 2.1-2.6, assumption 2.9 and assumption 3.1 hold with* $\delta = 4$. *Then,* $0 \leq \hat{\Delta}_n^Z \leq \hat{\Delta}_n^{ehw}$, *and* $\hat{\Delta}_n^Z \xrightarrow{p} \Delta^Z$, *where* $\Delta^{cond} \leq \Delta^Z \leq \Delta^{ehw}$, *and all inequalities are in a matrix sense.*

Then, I can estimate $V^{\text{causal}}$ by replacing $\Delta^{\text{cond}}$ with the estimate $\hat{\Delta}^Z$. That is, I propose the estimator:

$$\hat{V}^{\text{causal}} = \hat{\Gamma}_n^{-1}\left(\hat{\rho}_n\hat{\Delta}_n^Z + (1-\hat{\rho}_n)\hat{\Delta}_n^{\text{ehw}}\right)\hat{\Gamma}_n^{-1}$$

for $V^{\text{causal}}$. By lemma 4.2, this estimator is not larger than $\hat{V}^{\text{ehw}}$ and is often smaller while remaining conservative when $n$ is large enough.

# 5 Simulations

In this section I use simulations from a simple data-generating process to illustrate how the variance estimator proposed in this article can improve upon the conventional robust variance estimator.

I focus on a single causal variable $W_{n,i}$. I generate attributes $X_{n,i}$, which contains a constant vector of 1s and $k$ standard normal variables. Population values of $\xi_{n,i}$ are generated as independent draws from a standard normal distribution. A single instrument $Z_{n,i}$ is generated as an independent draw from a standard normal distribution. Then I generate $W_{n,i}$ as

$$W_{n,i} = \delta Z_{n,i} + \nu \xi_{n,i}$$

The parameter $\delta$ controls the strength of the instrument and $\nu$ controls the degree of endogeneity in $W_{n,i}$. This setup is consistent across all three simulations.

In the first set of simulations, I draw heterogeneous treatment effects for each observation from a normal distribution with mean depending on the attributes

$$\theta_{n,i} \sim \mathrm{N}(\psi' X_{n,i}, \sigma^2)$$

Then, I generate outcomes $Y_{n,i}$ as

$$Y_{n,i} = \theta_{n,i} W_{n,i} + \xi_{n,i}$$

$E[Z_{n,i}] = 0$, so $\Xi_n$ is also a row-vector of zeros and $V_{n,i} = Z_{n,i}$ Moreover, $\Gamma = \delta$ since

$$E[V_{n,i} U_{n,i}'] = \delta E[Z_{n,i}^2] + \nu E[Z_{n,i}] E[\xi_{n,i}] = \delta$$

which follows since $Z_{n,i}'$ is distributed $\chi_1^2$. I estimate $\hat{\theta}^{\text{causal}}$ using

In the second and third set of simulations presented in tables 2 and 3, I generate one attribute, so $k = 1$ across all specifications. This attribute, $X_{n,i}$ is demeaned. I generate heterogeneous treatment effects for each observation through an interaction structure:

$$\theta_{n,i} = 1 + \beta X_{n,i}.$$

I generate outcomes $Y_{n,i}$ as

$$Y_{n,i} = \theta_{n,i} W_{n,i} + \xi_{n,i}$$

$$Y_{n,i} = W_{n,i} + \beta W_{n,i} X_{n,i} + \xi_{n,i}.$$

Since $X_{n,i}$ has mean 0, $\theta_n^{\text{causal}} = \frac{1}{n}\sum_{i=1}^{n} \theta_{n,i} = 1$ and can be consistently estimated by either performing 2SLS using $Z_{n,i}$ as an instrument for $W_{n,i}$ alone, or through using 2SLS using $Z_{n,i}$ and $Z_{n,i}X_{n,i}$ as instrument for $W_{n,i}$ and $W_{n,i}X_{n,i}$ respectively. The former is examined in table 2; the latter is examined in table 3.

1. $(\hat{V}^{\text{ehw}}/N)^{1/2}$: the standard errors reported using the usual EHW robust standard errors.

2. $(\hat{V}^{\text{causal}}/N)^{1/2}$: the standard errors reported incorporating $\hat{\Delta}^{Z}$.

3. Coverage rate using $\hat{V}^{\text{causal}}$: The percentage of iterations in which $\theta_{\text{causal}}$ falls within the estimated 95 % confidence interval surrounding $\hat{\theta}^{\text{causal}}$ using EHW standard errors.

4. Coverage rate using $\hat{V}^{\text{caus}}$: The percentage of iterations in which $\theta^{\text{causal}}$ falls within the estimated 95 % confidence interval surrounding $\hat{\theta}^{\text{causal}}$ using the adjusted standard errors proposed in this paper.

Results varying the number of control variables $k$ are reported in Table 1. In all of these specifications, $\psi_1 = 0$ and $\psi_j = 2$ for all $j > 1$.

The first row across all tables, labeled MC Standard Errors, calculates the standard error of the estimates $\hat{\theta}$ across 10000 replications. The theoretical results of this paper demonstrate that $\bar{V}^{\text{Causal}}$ should converge to something larger than the true variance. Thus, $(\bar{V}^{\text{Causal}}/N)^{1/2}$ should converge to an object larger than the target estimate in the first row.

In table 1, $\hat{V}^{\text{Causal}}$ is lower than $\hat{V}^{\text{EHW}}$ across all specifications, as predicted by theory. When $n$ =1000, $k = 2, 3$ and 4 the confidence intervals undercover: the true parameter only falls within the estimated confidence intervals 94.5 %, 94.3 %, and 94.3 % of the iterations respectively. This is consistent with the results in Abadie et. al (2020) [1]. Theory suggests that $\hat{\Delta}^{Z}$ should converge to a conservative estimate, so I include two specifications, both with $n$ = 10000, and with $k = 3$ and $k = 10$ respectively. In these cases, the coverage rates are 95.4% and 95.3% , suggesting that as I increase $n$ the estimate is converging the way theory predicts. It's also interesting that in both these cases the EHW coverage rate is 96.6 %, so even when $n$ is large the EHW standard errors can overcover. This is also consistent with theoretical results demonstrating $\hat{\Delta}^{\text{ehw}}$ is conservative.

Tables 2 and 3 demonstrate how incorporating interaction effects can improve the performance of the traditional EHW variance estimator when interaction effects explain treatment effect heterogeneity. In table 2, interactions are not accounted for the EHW estimator is conservative across all specifications due to the presence of heterogeneous treatment effects generated by interaction effects. $\hat{V}^{\text{Caus}}$ is less conservative across all specifications.

In table 3, interaction effects are directly controlled for in the regression. Across all specifications, the traditional EHW variance estimator performs very well. This is because after controlling for interactions, there is no treatment effect heterogeneity and–as section 3.6 shows–in this case, this estimator is no longer conservative. $\hat{V}^{\text{Caus}}$ is nearly identical to the EHW variance estimator across all specifications. This suggests that when treatment effects are constant within the regression model, $\hat{V}^{\text{Caus}}$ still performs well, but is an unnecessary adjustment.

## 6 Conclusion

I show that the usual robust standard errors for inference on IV estimators is overly conservative when the sample makes up a non-negligible fraction of the population of interest.

In this paper I notice that the choice of 2SLS is not unique within this framework; for instance, a study of the GMM estimator and its properties would be a natural extension of this paper.

A natural extension of these ideas would be to derive an improved Hausman test in this setting. Moreover, this paper makes the strong assumption of randomized sampling. More complicated sampling schemes are natural to consider in this context, such as stratified sampling designs where some group of observations are more likely to be sampled than others.

While the estimator proposed in this paper improves on the usual robust standard error estimator currently commonly used in empirical research, there might still be room for improvement using other methods. For instance, instead of linear regression methods, non-linear methods might prove more effective when appropriate.

## 7 Tables

| $n$ | 1000 | 1000 | 1000 | 1000 | 1000 | 10000 |
|---|---|---|---|---|---|---|
| $\rho$ | .5 | .5 | .5 | .5 | .5 | .5 |
| $k$ | 1 | 2 | 3 | 4 | 10 | 3 |
| $\delta$ | 1 | 1 | 1 | 1 | 1 | 1 |
| MC Standard Deviation | 0.19 | 0.264 | 0.298 | 0.334 | 0.168 | 0.096 |
| $(\bar{V}^{\text{EHW}}/N)^{1/2}$ | 0.201 | 0.276 | 0.311 | 0.349 | 0.18 | 0.103 |
| $(\bar{V}^{\text{Caus}}/N)^{1/2}$ | 0.191 | 0.26 | 0.292 | 0.328 | 0.168 | 0.097 |
| $\frac{(\bar{V}^{\text{EHW}}/N)^{1/2}}{(\bar{V}^{\text{Caus}}/N)^{1/2}}$ | 0.951 | 0.942 | 0.94 | 0.94 | 0.938 | 0.942 |
| CR using $\bar{V}^{\text{EHW}}$ | 0.96 | 0.958 | 0.959 | 0.956 | 0.966 | 0.966 |
| CR using $\bar{V}^{\text{Caus}}$ | 0.95 | 0.945 | 0.943 | 0.943 | 0.954 | 0.953 |

Table 1: Simulation Results

| $n$ | 1000 | 1000 | 1000 | 1000 | 10000 | 10000 |
|---|---|---|---|---|---|---|
| $\rho$ | .5 | .5 | .5 | .5 | .5 | .5 |
| $\beta$ | 1 | 5 | 10 | 25 | 1 | 5 |
| $\delta$ | 1 | 1 | 1 | 1 | 1 | 1 |
| MC Standard Deviation | 0.096 | 0.426 | 0.835 | 2.1 | 0.03 | 0.133 |
| $(\bar{V}^{\text{EHW}}/N)^{1/2}$ | 0.099 | 0.447 | 0.895 | 2.191 | 0.032 | 0.141 |
| $(\bar{V}^{\text{Causal}}/N)^{1/2}$ | 0.094 | 0.417 | 0.835 | 2.049 | 0.03 | 0.132 |
| $\frac{(\bar{V}^{\text{EHW}}/N)^{1/2}}{(\bar{V}_{\text{Caus}}/N)^{1/2}}$ | 0.947 | 0.933 | 0.934 | 0.935 | 0.949 | 0.936 |
| CR using $\bar{V}^{\text{EHW}}$ | 0.956 | 0.959 | 0.957 | 0.96 | 0.963 | 0.962 |
| CR using $\bar{V}^{\text{Causal}}$ | 0.942 | 0.944 | 0.943 | 0.945 | 0.95 | 0.946 |

Table 2: Simulation Results Without Accounting for Interaction

| $n$ | 1000 | 1000 | 1000 | 1000 | 10000 | 10000 |
|---|---|---|---|---|---|---|
| $\rho$ | .5 | .5 | .5 | .5 | .5 | .5 |
| $\beta$ | 1 | 5 | 10 | 25 | 1 | 5 |
| $\delta$ | 1 | 1 | 1 | 1 | 1 | 1 |
| MC Standard Deviation | 0.046 | 0.045 | 0.045 | 0.045 | 0.014 | 0.014 |
| $(\bar{V}^{\text{EHW}}/N)^{1/2}$ | 0.046 | 0.044 | 0.045 | 0.045 | 0.014 | 0.014 |
| $(\bar{V}^{\text{Causal}}/N)^{1/2}$ | 0.046 | 0.044 | 0.045 | 0.045 | 0.014 | 0.014 |
| $\frac{(\bar{V}^{\text{EHW}}/N)^{1/2}}{(\bar{V}_{\text{Caus}}/N)^{1/2}}$ | 1 | 1 | 1 | 1 | 1 | 1 |
| CR using $\bar{V}^{\text{EHW}}$ | 0.946 | 0.946 | 0.952 | 0.951 | 0.952 | 0.951 |
| CR using $\bar{V}^{\text{Causal}}$ | 0.946 | 0.946 | 0.952 | 0.951 | 0.952 | 0.951 |

Table 3: Simulation Results Accounting for Interaction

# 8 Proofs

*Proof of Theorem 1.* For $n$ large enough $\sum_{i=1}^{n} X_{n,i}X_{n,i}^{'}$ is full rank and $\Lambda_n, \Xi_n$ exists.

$$\begin{aligned}
\sum_{i=1}^{n} E[W_{n,i}^{VX}] &= \sum_{i=1}^{n} E[Z_{n,i}X_{n,i}^{'} - \Xi_n \sum_{i=1}^{n} X_{n,i}X_{n,i}^{'}] \\
&= E[\sum_{i=1}^{n}(Z_{n,i}X_{n,i}^{'}) - \Big(\sum_{i=1}^{n} E[Z_{n,i}]X_{n,i}^{'}\Big)\Big(\sum_{i=1}^{n} X_{n,i}X_{n,i}^{'}\Big)^{-1}\Big(\sum_{i=1}^{n} X_{n,i}X_{n,i}^{'}\Big)] \\
&= \sum_{i=1}^{n} E[Z_{n,i}X_{n,i}^{'}] - E\Big[(\sum_{i=1}^{n} E[Z_{n,i}]X_{n,i}^{'})\Big] = \sum_{i=1}^{n} E[Z_{n,i}X_{n,i}^{'}] - \sum_{i=1}^{n} E[Z_{n,i}X_{n,i}^{'}]\Big) = 0
\end{aligned}$$

where $X_{n,i}$ can be folded inside the expectation since the attributes are nonstochastic. Thus, $\sum_{i=1}^{n} E[W_{n,i}^{VX}] = 0$.

For all $n$ large enough, $\Xi_n = C_n$, which implies $E[V_{n,i}] = 0$. Thus

$$E[V_{n,i}Y_{n,i}] = E[V_{n,i}, W_{n,i}]\theta_{n,i} + E[V_{n,i}]\xi_{n,i} = E[V_{n,i}W_{n,i}^{'}]\theta_{n,i}$$

**Lemma 8.1** (Analog to Lemma A.2 in Abadie et. al)**.**

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{n} R_{n,i}V_{n,i}\epsilon_{n,i} \xrightarrow{d} \mathcal{N}(0, \Delta^{cond} + (1-\rho)\Delta^{\mu})$$

*Proof.* Let $A_{n,i} = a^{'}X_{n,i}\epsilon_{n,i}$ for $a \in \mathbf{R}^k$. First, I show that $A_{n,i}$ satisfies the conditions of Lemma $A$.1 in Abadie et. al

$$\frac{1}{n}\sum_{i=1}^{n} E\Big[|A_{n,i}^{2+\delta}\Big] \leq \frac{||a||^{2+\delta}}{n}\sum_{i=1}^{n} E\Big[||V_{n,i}||^{2+\delta}(|Y_{n,i}| + ||W_{n,i}||\;||\theta_n|| + ||X_{n,i}||\;||\gamma_n||)^{2+\delta}\Big]$$

□

By Minkowski's inequality:

$$\begin{aligned}
\left(\frac{1}{n}\sum_{i=1}^{n} E\left[\Big(||V_{n,i}||(|Y_{n,i}| + ||U_{n,i}||\;||\theta_n|| + ||X_{n,i}||\;||\gamma_n||)\Big)^{2+\delta}\right]\right)^{\frac{1}{2+\delta}} \leq \frac{1}{n}\sum_{i=1}^{n} E\left[\Big(||V_{n,i}||\;|Y_{n,i}|\Big)^{2+\delta}\right]^{\frac{1}{2+\delta}} + \frac{1}{n}\sum_{i=1}^{n} E\left[\Big(||V_{n,i}||\;||W_{n,i}||\;||\theta_n||\Big)^{2+\delta}\right]^{\frac{1}{2+\delta}} \\
+ \frac{1}{n}\sum_{i=1}^{n} E\left[\Big(||V_{n,i}||\;||X_{n,i}||\;||\gamma_n||\Big)^{2+\delta}\right]^{\frac{1}{2+\delta}} \qquad (2)
\end{aligned}$$

By Cauchy-Schwarz,

$$\left(\frac{1}{n}\sum_{i=1}^{n} E\left[\left(||V_{n,i}||\,|Y_{n,i}|\right)^{2+\delta}\right]\right)^{2} \leq \frac{1}{n}\sum_{i=1}^{n} E\left[\left(||V_{n,i}||]\right)^{4+4\delta+\delta^2}\right]E\left[\left(|Y_{n,i}|]\right)^{4+4\delta+\delta^2}\right]$$

which is bounded since it is a product of bounded functions by Assumption 3.3. Note that this implies $\frac{1}{n}\sum_{i=1}^{n} E\left[\left(||V_{n,i}||\ |Y_{n,i}|\right)^{2+\delta}\right]$ is bounded. Identical arguments show the other elements of the sum on the left-hand side of equation 2 are bounded as well: thus, the left-hand side is bounded. Since $x^{\frac{1}{2+\delta}}$ is an increasing function for positive $x$, this also creates a bound for $\frac{1}{n}\sum_{i=1}^{n} E\left[|A_{n,i}^{2+\delta}\right]$

The following condition is also satisfied:

$$\sum_{i=1}^{n} \mu_{n,i} = a^{'} \sum_{i=1}^{n} E[V_{n,i}\epsilon_{n,i}] = 0.$$

Then,

$$a^{'}\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(V_{n,i}\epsilon_{n,i})\right)a \rightarrow a^{'}\Delta^{\mathrm{cond}}a > 0$$
$$a'\left(\frac{1}{n}\sum_{i=1}^{n} E\left[V_{n,i}\epsilon_{n,i}\right]E\left[\epsilon_{n,i}V_{n,i}^{'}\right]\right)a \rightarrow a^{'}\Delta^{\mu}a.$$

Applying Lemma A.1 in Abadie et. al, then

$$a^{'}\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{n} R_{n,i}V_{n,i}\epsilon_{n,i}\Big) \xrightarrow{d} \mathcal{N}\Big(0, a^{'}(\Delta^{\mathrm{cond}} + (1-\rho)\Delta^{\mu})a\Big).$$

Applying Cramer-Wold

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{n} R_{n,i}V_{n,i}\epsilon_{n,i} \xrightarrow{d} \mathcal{N}\Big(0, \Delta^{\mathrm{cond}} + (1-\rho)\Delta^{\mu}\Big)$$

□

*Proof of Theorem 3.2.* Suppose Assumptions 2.1-2.6, and 2.9 hold. Then,

$$\sqrt{N}(\hat{\theta} - \theta_n^{causal}) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1}(\rho\Delta^{\mathrm{cond}} + (1-\rho)\Delta^{\mathrm{ehw}})\Gamma^{-1})$$

For ease of notation I consider the case when $L = K$ and the model is just-identified. With probability approaching 1,

$$\sum_{i=1}^{n} R_{n,i} \begin{pmatrix} V_{n,i}W'_{n,i} & V_{n,i}X'_{n,i} \\ X_{n,i}V'_{n,i} & X_{n,i}X'_{n,i} \end{pmatrix}$$

is invertible. Then

$$\begin{aligned} \sqrt{N}\begin{pmatrix} \hat{\theta}_n - \theta_n^{causal} \\ \hat{\gamma}_n - \gamma_n^{causal} \end{pmatrix} &= \left(\frac{1}{N}\sum_{i=1}^{n} R_{n,i}\begin{pmatrix} V_{n,i}W'_{n,i} & V_{n,i}X'_{n,i} \\ X_{n,i}V_{n,i} & X_{n,i}X_{n,i} \end{pmatrix}\right)^{-1} \sum_{i=1}^{n} R_{n,i}\begin{pmatrix} V_{n,i}\epsilon_{n,i} \\ X_{n,i}\epsilon_{n,i} \end{pmatrix} \\ &= \begin{pmatrix} \Omega_n^{VW} & \Omega_n^{VX} \\ \Omega_n^{XV} & \Omega_n^{XX} \end{pmatrix}^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^{n} R_{n,i}\begin{pmatrix} V_{n,i}\epsilon_{n,i} \\ X_{n,i}\epsilon_{n,i} \end{pmatrix} + r_n. \end{aligned}$$

Where

$$r_n = \left[\begin{pmatrix} \tilde{P}_n^{VW} & \tilde{P}_n^{VX} \\ \tilde{P}_n^{XV} & \tilde{P}_n^{XX} \end{pmatrix}^{-1} - \begin{pmatrix} \Omega_n^{VU} & \Omega_n^{VX} \\ \Omega_n^{XV} & \Omega_n^{XX} \end{pmatrix}^{-1}\right] \frac{1}{\sqrt{N}}\sum_{i=1}^{n} R_{n,i}\begin{pmatrix} V_{n,i}\epsilon_{n,i} \\ X_{n,i}\epsilon_{n,i} \end{pmatrix}$$

$\Omega^{VX} = 0$, $\begin{pmatrix} \tilde{W}_n^{VW} & \tilde{W}_n^{VX} \\ \tilde{W}_n^{XV} & \tilde{W}_n^{XX} \end{pmatrix}^{-1} - \begin{pmatrix} \Omega_n^{VW} & \Omega_n^{VX} \\ \Omega_n^{XV} & \Omega_n^{XX} \end{pmatrix}^{-1}$ is $o_p(1)$, and $\frac{1}{\sqrt{N}}\sum_{i=1}^{n} R_{n,i}V_{n,i}, \epsilon_{n,i}$ is $O_p(1)$ by Lemma 3.1. Thus

$$\sqrt{N}(\hat{\theta}_n - \theta_n^{causal}) = (\Omega^{VW})^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{n} R_{n,i}V_{n,i}\epsilon_{n,i} + o_p(1).$$

if $\frac{1}{\sqrt{N}}\sum_{i=1}^{n} R_{n,i}X_{n,i}\epsilon_{n,i} = O_p(1)$. Following the approach in Abadie et al., rewrite this sum as

$$\left(\frac{n\rho_n}{N}\right)^{1/2}\left[\frac{1}{n^{\frac{1}{2}}}\sum_{i=1}^{n}\left(\frac{R_{n,i}}{\sqrt{\rho_n}}\right)X_{n,i}\epsilon_{n,i}\right]$$

$(\frac{n\rho_n}{N})^{\frac{1}{2}} \xrightarrow{p} 1$, so all I need to show is that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{R_{n,i}}{\sqrt{\rho_n}}\right)X_{n,i}\epsilon_{n,i} = O_p(1)$$

Notice that because the $X_{n,i}$ are nonrandom and $R_{n,i}$ is independent of $\epsilon_{n,i}$

$$\begin{aligned}
E\Bigg[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg(\frac{R_{n,i}}{\sqrt{\rho_n}}\Bigg)X_{n,i}\epsilon_{n,i} = O_p(1)\Bigg] &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg(\frac{E\Big[R_{n,i}\Big]}{\sqrt{\rho_n}}\Bigg)E\Big[X_{n,i}\epsilon_{n,i}\Big] \\
&= \sqrt{\frac{\rho_n}{n}}\sum_{i=1}^{n}E\Big[X_{n,i}\epsilon_{n,i}\Big] = 0
\end{aligned}$$

Now I show that the variances of each of these terms are bounded. Consider the $j$th element of this vector. Then by independence across $i$

$$\begin{aligned}
var\Big[\frac{1}{n^{\frac{1}{2}}}\sum_{i=1}^{n}(\frac{R_{n,i}}{\sqrt{\rho_n}})X_{n,i,j}\epsilon_{n,i})\Big] &= \frac{1}{n}\sum_{i=1}^{n}var\Big[(\frac{R_{n,i}}{\sqrt{\rho_n}}X_{n,i,j}\epsilon_{n,i}\Big] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}E\Big(\Big[\frac{R_{n,i}}{\sqrt{\rho_n}}X_{n,i,j}\epsilon_{n,i}\Big]\Big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}X_{n,i,j}^2E(\epsilon_{n,i,j}^2).
\end{aligned}$$

Since $R_{n,i}^2 = R_{n,i}$, $E(R_{n,i}) = \rho_n$ and $X_{n,i,j}$ is nonstochastic. $X_{n,i,j}^2$ and $E[\epsilon_{n,i}^2]$ are bounded by assumption 3.3. □

*Proof of Theorem 1.* Let $\hat{\Xi}_n = \left(\sum_{i=1}^{n} R_{n,i}Z_{n,i}X_{n,i}'\right)(R_{n,i}X_{n,i}X_{n,i})$

Note that with probability approaching one, $\Xi_n$ exist and is equal to $C_n$. Then,

$$\begin{aligned}
\hat{\Xi}_n - \Xi_n &= \left(\frac{1}{N}\sum_{i=1}^{n}R_{n,i}Z_{n,i}X_{n,i}'\right)\left(\frac{1}{N}\sum_{i=1}^{n}R_{n,i}X_{n,i}X_{n,i}'\right)^{-1} - C_n \\
&= \left(\frac{1}{N}\sum_{i=1}^{n}R_{n,i}Z_{n,i}X_{n,i}'\right)\left(\frac{1}{N}\sum_{i=1}^{n}R_{n,i}X_{n,i}X_{n,i}'\right)^{-1} - C_n\left(\frac{1}{N}\sum_{i=1}^{n}R_{n,i}X_{n,i}X_{n,i}'\right)\left(\frac{1}{N}\sum_{i=1}^{n}R_{n,i}X_{n,i}X_{n,i}'\right)^{-1} \\
&= \left(\frac{1}{N}R_{n,i}\Big[Z_{n,i} - C_nX_{n,i}\Big]X_{n,i}'\right)\left(\frac{1}{N}\sum_{i=1}^{n}R_{n,i}X_{n,i}X_{n,i}'\right)^{-1} \\
&= \left(\frac{1}{N}\sum_{i=1}^{n}R_{n,i}V_{n,i}X_{n,i}'\right)\left(\frac{1}{N}\sum_{i=1}^{n}R_{n,i}X_{n,i}X_{n,i}'\right)^{-1}
\end{aligned}$$

which converges to 0.

Let $\hat{P}_n^{WV} = \frac{1}{N}\sum_{i=1}^n R_{n,i}W_{n,i}\left(Z_{n,i} - \hat{\Xi}_n X_{n,i}\right)^{'}$. Then,

$$\begin{aligned}\Omega_n^{WV} - \tilde{P}_n^{WV} &= \frac{1}{N}\sum_{i=1}^n R_{n,i}\left(W_{n,i}\Big[Z_{n,i} - \Xi_n X_{n,i}\Big]^{'} - W_{n,i}\Big[Z_{n,i} - \hat{\Xi}_n X_{n,i}\Big]^{'}\right)\\ &= \frac{1}{N}\sum_{i=1}^n R_{n,i}\left(W_{n,i}\Big[(\hat{\Xi}_n - \Xi_n)X_{n,i}\Big]^{'}\right) = \tilde{P}^{WX}(\hat{\Xi}_n - \Xi_n) \rightarrow 0\end{aligned}$$

.

This also implies that $\tilde{P}_n^{VW} - \Omega_n^{VW} \rightarrow 0$.

Now define $\tilde{P}_n^{VV} = \frac{1}{N}\sum_{i=1}^n R_{n,i}\left(Z_{n,i} - \hat{\Xi}_n X_{n,i}\right)\left(Z_{n,i} - \hat{\Xi}_n X_{n,i}\right)^{'}$. Then

$$\begin{aligned}\Omega_n^{VV} - \tilde{P}_n^{VV} &= \frac{1}{N}\sum_{i=1}^n R_{n,i}\left(\Big[Z_{n,i} - \Xi_n X_{n,i}\Big]\Big[Z_{n,i} - \Xi_n X_{n,i}\Big]^{'} - [Z_{n,i} - \hat{\Xi}_n X_{n,i}\Big]\Big[Z_{n,i} - \hat{\Xi}_n X_{n,i}\Big]^{'}\right)\\ &= (\hat{\Xi}_n - \Xi_n)\tilde{P}_n^{XX}(\hat{\Xi}_n - \Xi_n)^{'} - \tilde{P}^{XZ}(\hat{\Xi}_n - \Xi_n)^{'} - (\hat{\Xi}_n - \Xi_n)\tilde{P}^{XZ} \rightarrow 0\end{aligned}$$

This implies $\hat{\Gamma}_n \rightarrow \Gamma_n$.

Now consider three different objects

$$\begin{aligned}\check{\Delta}_n^{\text{ehw}} &= \frac{1}{N}\sum_{i=1}^n R_{n,i}V_{n,i}\hat{\epsilon}_{n,i}^2 V_{n,i}^{'}\\ \tilde{\Delta}_n^{\text{ehw}} &= \frac{1}{N}\sum_{i=1}^n R_{n,i}V_{n,i}\epsilon_{n,i}^2 V_{n,i}^{'}\\ \Delta_n^{\text{ehw}} &= \frac{1}{n}\sum_{i=1}^n V_{n,i}E[\epsilon_{n,i}^2]V_{n,i}^{'}\end{aligned}$$

By direct calculation, $\tilde{\Delta}^{\text{ehw}}$ can be written as a sum of elements of the following matrix, with some elements multiplied by the fixed causal estimands $\theta_n^{\text{causal}}$ and $\gamma_n^{\text{causal}}$:

$$\frac{1}{n}\sum_{i=1}^n R_{n,i}\begin{pmatrix} Y_{n,i}\\ W_{n,i}\\ V_{n,i}\\ X_{n,i}\end{pmatrix}\begin{pmatrix} Y_{n,i}\\ W_{n,i}\\ V_{n,i}\\ X_{n,i}\end{pmatrix}^{'} \bigotimes \begin{pmatrix} Y_{n,i}\\ W_{n,i}\\ V_{n,i}\\ X_{n,i}\end{pmatrix}\begin{pmatrix} Y_{n,i}\\ W_{n,i}\\ V_{n,i}\\ X_{n,i}\end{pmatrix}^{'}.$$

When $\delta = 4$ in 3.3

$$\frac{1}{n}\sum_{i=1}^{n} R_{n,i} \begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix} \begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix}^{'} \bigotimes \begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix} \begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix}^{'} \to \frac{1}{n}\sum_{i=1}^{n} E\left[ \begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix} \begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix}^{'} \bigotimes \begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix} \begin{pmatrix} Y_{n,i} \\ W_{n,i} \\ V_{n,i} \\ X_{n,i} \end{pmatrix}^{'} \right].$$

It follows directly that $\tilde{\Delta}_n^{\text{ehw}} \to \Delta_n^{\text{ehw}}$. By the convergence of $\hat{\theta}_n^{\text{causal}}$ and $\hat{\gamma}_n^{\text{causal}}$, $\check{\Delta}_n^{\text{ehw}} \to \tilde{\Delta}_n^{\text{ehw}}$. By the convergence of $Z_{n,i} - \hat{\xi}_n X_{n,i}$, $\hat{\Delta}_n^{\text{ehw}} \to \check{\Delta}_n^{\text{ehw}}$. Then,

$$\hat{\Delta}_n^{\text{ehw}} - \Delta^{\text{ehw}} = (\hat{\Delta}_n^{\text{ehw}} - \check{\Delta}_n^{\text{ehw}}) + (\check{\Delta}_n^{\text{ehw}} - \tilde{\Delta}_n^{\text{ehw}}) + (\tilde{\Delta}_n^{\text{ehw}} - \Delta_n^{\text{ehw}}) + (\Delta_n^{\text{ehw}} - \Delta_n^{\text{ehw}}) \to 0$$

□

*Proof of Lemma 4.2.*

$$\hat{\Delta}^{\text{Z}} = \hat{\Delta}_n^{\text{ehw}} - \hat{\Delta}_n^{\text{proj}}, \;\; \hat{\Delta}_n^{\text{proj}} = \frac{1}{n}\sum_{i=1}^{n} R_{n,i}\hat{G}_n X_{n,i} X_{n,i}^{'} G_n^{'}$$

which implies that $\hat{\Delta}_n^{\text{ehw}} - \hat{\Delta}^{\text{Z}}$ is positive semi-definite.

Define

$$G_n = \left(\frac{1}{n}\sum_{i=1}^{n} E[V_{n,i}\epsilon_{n,i}]X_{n,i}^{'}\right)\left(X_{n,i}X_{n,i}^{'}\right)^{-1}, \;\; \Delta_n^{proj} = \frac{1}{n}\sum_{i=1}^{n} G_n X_{n,i} X_{n,i} G_n$$

By Lemma 2, $\hat{G}_n \to G_n$ and $\hat{\Delta}_n^{\text{proj}} \to \Delta_n^{\text{proj}}$. If $\Delta_n^{\text{ehw}} = \frac{1}{n}\sum_{i=1}^{n} E[V_{n,i}\epsilon_{n,i}^2 V_{n,i}^{'}]$ this in turn implies that $\hat{\Delta}_n^{\text{z}} \to \Delta_n^{\text{z}}$ where

$$\Delta_n^{\text{z}} = \Delta_n^{\text{ehw}} - \Delta_n^{\text{proj}}.$$

.

Let $\Delta_n^{\mu} = \frac{1}{n}\sum_{i=1}^{n} E[V_{n,i}\epsilon_{n,i}]E[V_{n,i\epsilon_{n,i}}\epsilon_{n,i}]^{'}$. Then

$$\Delta_n^{\mu} - \Delta_n^{\text{proj}} = \frac{1}{n}\sum_{i=1}^{n} E[V_{n,i}\epsilon_{n,i}]E[V_{n,i}\epsilon_{n,i}]^{'} - \left(\frac{1}{n}\sum_{i=1}^{n} E[V_{n,i}\epsilon_{n,i}]X_{n,i}^{'}\right)\left(X_{n,i}X_{n,i}^{'}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_{n,i}E[V_{n,i}\epsilon_{n,i}]^{'}\right)$$

Let $A_n$ and $D_n$ be the matrices with rows equal to $\frac{E[\epsilon_{n,i}V_{n,i}^{'}]}{\sqrt{n}}$ and $\frac{X_{n,i}^{'}}{\sqrt{n}}$ respectively. Let $I_n$ be the identity matrix of size $n$. Then

$$\Delta_n^{\mu} - \Delta_n^{\text{proj}} = A_n'(I_n - D_n(D_n' D_n)D_n')A_n$$

which is positive semidefinite. Note also that

$$\Delta_n^{\text{cond}} = \Delta_n^{\text{ehw}} - \Delta_n^{\mu} = (\Delta_n^{\text{ehw}} - \Delta_n^{\text{proj}}) - (\Delta_n^{\mu} - \Delta_n^{\text{proj}}) \leq \Delta_n^{\text{ehw}} - \Delta_n^{\text{proj}} = \Delta_n^{\text{Z}}.$$

Thus, $\Delta_n^{\text{ehw}} \geq \Delta_n^{\text{z}} \geq \Delta_n^{\text{cond}}$. Under the assumption that all of these matrices have limits,

$$\Delta_n^{\mu} \leq \Delta_n^{\text{z}} \leq \Delta_n^{\text{ehw}}$$

as desired. □

# References

[1] Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296, 2020.

[2] Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2023.

[3] Joshua D Angrist and Guido W Imbens. 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, 1995.

[4] David Card. Using geographic variation in college proximity to estimate the return to schooling, 1993.

[5] Charles F Manski and John V Pepper. How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics*, 100(2):232–244, 2018.

[6] Paul R Rosenbaum and Paul R Rosenbaum. *Overt bias in observational studies.* Springer, 2002.

[7] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.