Sampling-based *vs.* Design-based Uncertainty in Regression Analysis *

Alberto Abadie[†] Susan Athey[‡] Guido W. Imbens[§]
Jeffrey M. Wooldridge[¶]

Current version June 2017 – First version September 2013

Abstract

Consider a researcher estimating the parameters of a regression function based on data for all 50 states in the United States or on data for all visits to a website. What is the interpretation of the estimated parameters and the standard errors? In practice, researchers typically assume that the sample is randomly drawn from a large population of interest and report standard errors that are designed to capture sampling variation. This is common practice, even in applications where it is difficult to articulate what that population of interest is, and how it differs from the sample. In this article, we explore an alternative approach to inference, which is partly design-based. In a design-based setting, the values of some of the regressors can be manipulated, perhaps through a policy intervention. Design-based uncertainty emanates from lack of knowledge about the values that the regression outcome would have taken under alternative interventions. We derive standard errors that account for design-based uncertainty instead of, or in addition to, sampling-based uncertainty. We show that our standard errors in general are smaller than the infinite-population sampling-based standard errors and provide conditions under which they coincide.

*We are grateful for comments by Daron Acemoglu, Joshua Angrist, Matias Cattaneo, Jim Poterba, Tymon Słoczyński, Bas Werker, and seminar participants at Microsoft Research, Michigan, Brown University, MIT, Stanford, Princeton, NYU, Columbia, Tilburg University, the Tinbergen Institute, American University, Montreal, Michigan State, Maryland, Pompeu Fabra, Carlos III and University College London, three referees, a co-editor, and especially for discussions with Gary Chamberlain. An earlier version of this paper circulated under the title "Finite Population Causal Standard Errors" (Abadie et al. (2014)).

[†]Professor of Economics, Massachusetts Institute of Technology, and NBER, abadie@mit.edu.

[‡]Professor of Economics, Graduate School of Business, Stanford University, and NBER, athey@stanford.edu.

[§]Professor of Economics, Graduate School of Business, and department of Economics, Stanford University, and NBER, imbens@stanford.edu.

[¶]University Distinguished Professor, Department of Economics, Michigan State University, wooldri1@msu.edu

1 Introduction

The dominant approach to inference in regression analysis in the social sciences takes a sampling perspective on uncertainty. This perspective relies on the assumption that the observed units can be viewed as a sample drawn randomly from a large population of interest. In many cases this random sampling perspective is a natural and attractive one. For example, if one analyzes individual-level data from the U.S. Current Population Survey, the Panel Study of Income Dynamics, or the 1% public use sample from the U.S. Census, it is natural to regard the sample as a small random subset of the population of interest. In many other settings, however, this sampling perspective is less attractive. For example, suppose that the data set to be analyzed contains information on all 50 states of the United States, all the countries in the world, or all visits to a website. If, for all units in this data set, we observe an outcome and some attributes at a single point in time, and we ask how the average outcome varies across two subpopulations defined by these attributes, the answer is a quantity that is known with certainty. Hence, the standard error should be zero. However, researchers analyzing this type of data typically report standard errors that are formally justified by the random sampling perspective. This widespread practice implicitly forces the object of interest to be a data generating process, or superpopulation, from which the actual population is drawn at random. In such a setting, uncertainty arises from lack of observability of the superpopulation. While this may be an appealing framework in some instances, it is clearly not so in cases where the interest resides in an actual finite population and, in any event, a researcher may want to first define the object of interest and then use an appropriate mode of inference, rather than allowing the mode of inference to implicitly define the object of interest of her/his investigation.

In this article, we provide an alternative framework for the interpretation of uncertainty in regression analysis regardless of whether a fraction of the population or the entire population is included in the sample. While our framework accommodates sampling-based uncertainty, it also takes into account design-based uncertainty, which arises when the parameter of interest is defined in terms of the unobserved outcomes that some units would attain under a certain intervention. Design-based uncertainty is often explicitly accounted for in the analysis of randomized experiments where it is the basis of randomization inference (Neyman, 1923; Rosenbaum, 2002; Imbens and Rubin, 2015), but it is rarely explicitly acknowledged in regression analyses or, more generally, in observational studies (exceptions in special cases include Samii and Aronow, 2012;

Freedman, 2008a and 2008b; Lin, 2013).

To illustrate the differences between sampling-based inference and design-based inference, we present two examples in Tables 1 and 2. In the example of Table 1, there is a finite population consiting of n units, each characterized by the values of a pair of variables, Y_i and Z_i. Here, we can define an estimand as a function of the pairs $\{(Y_i, Z_i)\}_{i=1}^n$ for the entire population. For example, the estimand could be the difference in the population average value of the outcome, Y_i, by values of the attribute, Z_i. Uncertainty about the estimand exists when we observe the values (Y_i, Z_i) only for a subset of the population, the sample. In Table 1 inclusion of unit i in a sample is coded by the binary variable R_i. In this setting, an estimator can be naturally defined as the difference in the average value of the outcome, Y_i, by values of the attribute, Z_i, in the sample. Sampling-based inference uses information about the process that determines $R_1, \ldots, R_n$, to assess the variability of estimators across different potential samples.

Table 1: Sampling-based Uncertainty (✓ is observed, ? is missing)

Unit	Actual Sample			Alternative Sample I			Alternative Sample II			...
	Y_i	Z_i	R_i	Y_i	Z_i	R_i	Y_i	Z_i	R_i	...
1	✓	✓	1	?	?	0	?	?	0	...
2	?	?	0	?	?	0	?	?	0	...
3	?	?	0	✓	✓	1	✓	✓	1	...
4	?	?	0	✓	✓	1	?	?	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
n	✓	✓	1	?	?	0	?	?	0	...

Table 2 depicts a different scenario. We encounter again a finite population of size n. For each population unit we now observe the value of one of two variables, either $Y_i(1)$ or $Y_i(0)$, but not both. $Y_i(1)$ and $Y_i(0)$ represent the potential outcomes that unit i would attain under exposure or lack of exposure to certain intervention (or treatment) of interest. In Table 2 exposure to the intervention is coded by the binary treatment variable, X_i. We observe $Y_i(1)$ if $X_i = 1$, and $Y_i(0)$ if $X_i = 0$. The estimand is a function of the full set of pairs $\{(Y_i(1), Y_i(0))\}_{i=1}^n$, for example, the average causal effect $(1/n)\sum_{i=1}^n (Y_i(1) - Y_i(0))$. As in the first example, the estimator is a function of the observed data, e.g., the difference in the average of observed values of $Y_i(1)$ and

$Y_i(0)$. Design-based inference uses information about the process that determines $X_1, \ldots, X_n$, to assess the variability of estimators across different potential samples. Notice that, under this mode of inference, uncertainty about the estimand remains even when we observe the entire population, as in Table 2.

Table 2: Design-based Uncertainty (✓ is observed, ? is missing)

Unit	Actual Sample			Alternative Sample I			Alternative Sample II			... ...
	$Y_i(1)$	$Y_i(0)$	X_i	$Y_i(1)$	$Y_i(0)$	X_i	$Y_i(1)$	$Y_i(0)$	X_i	...
1	✓	?	1	✓	?	1	?	✓	0	...
2	?	✓	0	?	✓	0	?	✓	0	...
3	?	✓	0	✓	?	1	✓	?	1	...
4	?	✓	0	?	✓	0	✓	?	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
n	✓	?	1	?	✓	0	?	✓	0	...

More generally, of course, we can have complex missing data processes that combine features of these two examples, with some units not included in the sample at all, and with one of the two potential outcomes not observed for the sample units. The inferential procedures proposed in this article address both sources of variability. As the examples in Tables 1 and 2 illustrate, articulating the exact nature of the estimand of interest and the source of uncertainty that makes the estimator stochastic are crucial steps to valid inference. For this purpose, it will be useful to distinguish between descriptive estimands, where uncertainty stems solely from not observing all units in the population of interest, and causal estimands, where the uncertainty stems, at least partially, from unobservability of potential outcomes.

The main formal contribution of this article is to generalize the results for the approximate variance for multiple linear regression estimators associated with the work by Eicker (1967), Huber (1967), and White (1980a,b, 1982), EHW from hereon, in two directions. First, we allow sampling from a finite population and, second, we allow for design-based uncertainty in addition to, or instead of, the sampling-based uncertainty that the EHW results based on. The first generalization decreases the variance, and the second increases the variance. Incorporating these generalizations requires developing a new framework for regression analysis with assumptions

that differ from the standard ones. This framework nests as special cases the Neyman (1923), Samii and Aronow (2012), Freedman (2008a), Freedman (2008b), and Lin (2013) regression analyses for data from randomized experiments. We show that in large samples the widely used EHW robust standard errors are conservative. Moreover, we show that the presence of attributes – that is, immutable characteristics of the units – can be exploited to improve on the EHW variance estimator, and we propose variance estimators that do so. Finally, we show that in some special cases, in particular the case where the regression function is correctly specified, the EHW standard errors are asymptotically correct.

One important practical advantage of our framework is that it justifies non-zero standard errors in cases where we observe all units in the population but design-based uncertainty remains. A second advantage of the formal separation into sampling-based and design-based uncertainty is that it allows us to discuss the distinction between internal and external validity (Shadish et al., 2002; Manski, 2013; Deaton, 2010) in terms of these two sources of uncertainty. For internal validity there are no assumptions required on the sampling process, and conversely, for external validity there are no assumptions required on the design.

2 A Simple Example

In this section we set the stage for the problems discussed in the current article by discussing least squares estimation in a simple example with a single binary regressor. We make four points. First, we show how design-based uncertainty affects the variance of regression estimators. Second, we show that the standard Eicker-Huber-White (EHW) variance estimator remains conservative when we take into account design-based uncertainty. Third, we show that there is a simple finite-population correction to the EHW variance estimator for descriptive estimands but not for causal estimands. Fourth, we discuss the relation between the two sources of uncertainty and the notions of internal and external validity of the estimand.

We focus on a setting with a finite population of size n. We sample N units from this population, with $R_i \in \{0,1\}$ indicating whether a unit was sampled ($R_i = 1$) or not ($R_i = 0$), so that $N = \sum_{i=1}^{n} R_i$. There is a single binary regressor, $X_i \in \{0,1\}$, and n_x (resp. N_x) are the number of units in the population (resp. the sample) with $X_i = x$. To make the discussion specific, suppose the binary regressor X_i is an indicator for a state regulation, say the state having a minimum wage higher than the federal minimum wage. We view the regressor not as

a fixed attribute or characteristic of each unit, but instead as a cause or policy variable whose value could have been different from the observed value. This generates missing data of the type shown in Table 2, where only some of the states of the world are observed, implying that there is design-based uncertainty. Formally, using the Rubin causal model or potential outcome framework (Neyman, 1923; Rubin, 1974; Holland, 1986; Imbens and Rubin, 2015), we postulate the existence of two potential outcomes for each unit, denoted by $Y_i(1)$ and $Y_i(0)$, for state average earnings without and with a state minimum wage, with Y_i, the realized outcome, given the actual or prevailing minimum wage, defined as:

$$Y_i = Y_i(X_i) = \begin{cases} Y_i(1) & \text{if } X_i = 1, \\ Y_i(0) & \text{if } X_i = 0. \end{cases}$$

These potential outcomes are viewed as non-stochastic attributes for unit i, irrespective of the realized value of X_i. They, as well as the additional observed attributes, Z_i, remain fixed in repeated sampling thought experiments, whereas R_i and X_i are stochastic and, as a result, so are the realize outcomes, Y_i. In the current section we abstract from the presence of fixed observed attributes, Z_i, which will play an important role in Section 3. Let $\boldsymbol{Y}$, $\boldsymbol{Y}(1)$, $\boldsymbol{Y}(0)$, $\boldsymbol{R}$, and $\boldsymbol{X}$ be the n-vectors with i-th element equal to Y_i, $Y_i(1)$, $Y_i(0)$, R_i, and X_i respectively. For sampled units (units with $R_i = 1$) we observe X_i, and Y_i.

In general, estimands are functions of the full set of population values $(\boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{X}, \boldsymbol{R})$. We consider two types of estimands, descriptive and causal. If an estimand can be written as a function of $(\boldsymbol{Y}, \boldsymbol{X})$, free of dependence on $\boldsymbol{R}$ and on the potential outcomes beyond the realized outcome, we label it a *descriptive* estimand. Intuitively a descriptive estimand is an estimand whose value would be known with certainty if we observe all the realized values of all variables for all units in the population. If an estimand cannot be written as a function of $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{R})$ because it depends on the potential outcomes $\boldsymbol{Y}(1)$ and $\boldsymbol{Y}(0)$, then we label it a *causal* estimand.

We now consider in our binary regressor example three closely related estimands, one descriptive and two causal. The first estimand is the difference in population averages by the prevailing minimum wage,

$$\theta^{\text{descr}} = \theta^{\text{descr}}(\boldsymbol{Y}, \boldsymbol{X}) = \frac{1}{n_1} \sum_{i=1}^{n} X_i Y_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - X_i) Y_i.$$

This estimand is a function of $(\boldsymbol{Y}, \boldsymbol{X})$ and so it is a descriptive estimand. The second estimand

is the sample average causal effect,

$$\theta^{\text{causal,sample}} = \theta^{\text{causal,sample}}(\boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{R}) = \frac{1}{N} \sum_{i=1}^{n} R_i \Big(Y_i(1) - Y_i(0) \Big).$$

This estimand is a causal estimand: it cannot be written as a function of $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{R})$ because it depends on the potential outcomes $\boldsymbol{Y}(1)$ and $\boldsymbol{Y}(0)$. The third estimand is the population version of $\theta^{\text{causal,sample}}$:

$$\theta^{\text{causal}} = \theta^{\text{causal}}(\boldsymbol{Y}(1), \boldsymbol{Y}(0)) = \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i(1) - Y_i(0) \Big).$$

This is again a causal estimand.

Now let us turn to estimators. In general an estimator is a function of the values of Y_i, X_i and Z_i for the units in the sample, that is for the units with $R_i = 1$. We focus on the properties of a particular estimator:

$$\widehat{\theta} = \frac{1}{N_1} \sum_{i=1}^{n} R_i X_i Y_i - \frac{1}{N_0} \sum_{i=1}^{n} R_i (1 - X_i) Y_i,$$

which can be interpreted as a least squares estimator of the coefficient on X_i for the regression of Y_i on X_i and a constant. There are two sources of randomness in this estimator: a sampling component arising from the randomness of $\boldsymbol{R}$ and a design component arising from the randomness of $\boldsymbol{X}$. We refer to the uncertainty generated by the randomness in the sampling component as *sampling-based uncertainty*, and the uncertainty generated by the design component as *design-based uncertainty*.

We will next consider the the first two moments of $\widehat{\theta}$ under combinations of two assumptions. The first assumption is on the sampling mechanism.

Assumption 1. (Random Sampling / External Validity)

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

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

The second assumption is on the assignment mechanism.

Assumption 2. (Random Assignment / Internal Validity)

$$\Pr(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{R}) = 1 \Big/ \binom{n}{n_1},$$

for all n-vectors $\boldsymbol{x}$ *with* $\sum_{i=1}^{n} X_i = n_1$.

We start by studying the first moment of the estimator, conditional on (N_1, N_0), and only for the cases where $N_1 \geq 1$ and $N_0 \geq 1$ (and thus $n_1 \geq 1$ and $n_0 \geq 1$). We leave this latter conditioning implicit in the notation throughout this section. A supplementary appendix contains proofs of the results in this section. First, taking the expectation only over the random sampling, under Assumption 1:

$$E\big[\widehat{\theta} \,|\, \boldsymbol{X}, N_1, N_0\big] = \theta^{\text{descr}}. \tag{2.1}$$

Notice that this result does not require random assignment. Second, taking the expectation only over the random assignment, under Assumption 2:

$$E\big[\widehat{\theta} \,|\, \boldsymbol{R}, N_1, N_0\big] = \theta^{\text{causal,sample}}. \tag{2.2}$$

This equality does not require random sampling. Third, taking the expectation over both the sampling and the assignment, maintaining both Assumptions 1 and 2:

$$E\big[\widehat{\theta} \,|\, N_1, N_0\big] = E\big[\theta^{\text{descr}} \,|\, N_1, N_0\big] = E\big[\theta^{\text{causal,sample}} \,|\, N_1, N_0\big] = \theta^{\text{causal}}.$$

Next we look at the variance of the estimator. Here we maintain both the random assignment and random sampling assumption. From Equations (2.1) and (2.2), it follows that $\text{var}(\widehat{\theta} \,|\, \boldsymbol{X}, N_1, N_0)$ measures dispersion with respect to θ^{descr}, while $\text{var}(\widehat{\theta} \,|\, \boldsymbol{R}, N_1, N_0)$ measures dispersion with respect to $\theta^{\text{causal,sample}}$. By the law of total variance, we can decompose:

$$\begin{aligned}
\text{var}\big(\widehat{\theta} \,|\, N_1, N_0\big) &= E\left[\text{var}\big(\widehat{\theta} \,|\, \boldsymbol{X}, N_1, N_0\big) \,\big|\, N_1, N_0\right] + \text{var}\big(\theta^{\text{descr}} \,|\, N_1, N_0\big) \\
&= E\left[\text{var}\big(\widehat{\theta} \,|\, \boldsymbol{R}, N_1, N_0\big) \,|\, N_1, N_0\right] + \text{var}\big(\theta^{\text{causal,sample}} \,|\, N_1, N_0\big).
\end{aligned} \tag{2.3}$$

Let

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i(1) - \frac{1}{n} \sum_{j=1}^{n} Y_j(1) \right)^2.$$

S_0^2 and S_θ^2 are analogously defined for $Y_1(0), \ldots, Y_n(0)$ and $\theta_1, \ldots, \theta_n$, respectively, where $\theta_i = Y_i(1) - Y_i(0)$. The variance of $\widehat{\theta}$ can be expressed as

$$\begin{aligned}
V^{\text{total}}(N_1, N_0, n_1, n_0) &= \text{var}\big(\widehat{\theta} \,|\, N_1, N_0\big), \\
&= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n},
\end{aligned} \tag{2.4}$$

which is a variant of the result in Neyman (1923).

For the first decomposition in equation (2.3), the sampling-based component of the total variance is

$$\begin{aligned} V^{\text{sampling}}(N_1, N_0, n_1, n_0) &= E\left[\text{var}\big(\widehat{\theta} \,|\, \boldsymbol{X}, N_1, N_0\big) \,\big|\, N_1, N_0\right], \\ &= \frac{S_1^2}{N_1}\left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0}\left(1 - \frac{N_0}{n_0}\right), \end{aligned}$$

and the design-based component, beyond the sampling-based component, is

$$\begin{aligned} V^{\text{design|sampling}}(N_1, N_0, n_1, n_0) &= \text{var}\big(\theta^{\text{descr}} \,|\, N_1, N_0\big) \\ &= \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\theta^2}{n}. \end{aligned}$$

For the second decomposition in equation (2.3), the design-based component of the variance is

$$\begin{aligned} V^{\text{design}}(N_1, N_0, n_1, n_0) &= E\left[\text{var}\big(\widehat{\theta} \,|\, \boldsymbol{R}, N_1, N_0\big) \,|\, N_1, N_0\right] \\ &= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N}, \end{aligned}$$

and the sample-based component, beyond the design-based component, is

$$\begin{aligned} V^{\text{sampling|design}}(N_1, N_0, n_1, n_0) &= \text{var}\big(\theta^{\text{causal,sample}} \,|\, N_1, N_0\big) \\ &= \frac{S_\theta^2}{N}\left(1 - \frac{N}{n}\right). \end{aligned}$$

Comment 1. Causal versus Descriptive Estimands

A key comparison is between the sampling variance for the estimator for the descriptive estimand and the design variance for the estimator for the sample average causal effect

$$V^{\text{sampling}}(N_1, N_0, n_1, n_0) = \frac{S_1^2}{N_1}\left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0}\left(1 - \frac{N_0}{n_0}\right),$$

versus

$$V^{\text{design}}(N_1, N_0, n_1, n_0) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N}.$$

In general these variances cannot be ranked: the sampling variance can be very close to zero if the sampling rate N/n is close to one, but it can also be larger than the design variance if the sampling rate is small and the variance of the treatment effect, S_θ^2, is substantial. □

Comment 2. Finite Population Correction

If the estimand is θ^{causal} or θ^{descr}, ignoring the fact that the population is finite generally leads to an overstatement of the variance:

$$V^{\text{total}}(N_1, N_0, \infty, \infty) - V^{\text{total}}(N_1, N_0, n_1, n_0) = \frac{S^2_\theta}{n} \geq 0,$$

$$V^{\text{sampling}}(N_1, N_0, \infty, \infty) - V^{\text{sampling}}(N_1, N_0, n_1, n_0) = \frac{S^2_1}{n_1} + \frac{S^2_0}{n_0} \geq 0.$$

If the estimand is $\theta^{\text{causal,sample}}$, however, the population size is irrelevant:

$$V^{\text{design}}(N_1, N_0, \infty, \infty) = V^{\text{design}}(N_1, N_0, n_1, n_0). \qquad \square$$

Comment 3. Large Population *vs* Sample is identical to Population

If the population is large relative to the sample, the incremental design-based component is zero, and the sampling-based variance component is equal to the total variance:

$$V^{\text{design|sampling}}(N_1, N_0, \infty, \infty) = 0, \quad V^{\text{sampling}}(N_1, N_0, \infty, \infty) = V^{\text{total}}(N_1, N_0, \infty, \infty).$$

In other words, if the population is large relative to the sample, it is sufficient to consider the sampling-based variance. This can be viewed as the implicit justification for the common practice of ignoring design-based uncertainty. If, at the other extreme, the sample is equal to the population, the sampling-based variance component is zero and the design-based component is equal to the total variance:

$$V^{\text{sampling}}(N_1, N_0, N_1, N_0) = 0, \quad V^{\text{design|sampling}}(N_1, N_0, N_1, N_0) = V^{\text{total}}(N_1, N_0, N_1, N_0).$$

$\square$

Comment 4. Internal versus External Validity

Often researchers are concerned about both the internal and external validity of estimands and estimators (Shadish et al., 2002; Manski, 2013; Deaton, 2010). The distinction between sampling and design-based uncertainty allows us to clarify these concerns. Internal validity bears on the question of whether $\widehat{\theta}$ is a good estimator for $\theta^{\text{causal,sample}}$. This relies on random assignment of the treatment. Whether or not the sampling is random is irrelevant for this question because $\theta^{\text{causal,sampling}}$ conditions on which units were sampled. External validity bears on the question whether $\widehat{\theta}$ is a good estimator for θ^{descr}. This relies on the random sampling assumption and

is not affected by the assumptions on the assignment process. However, for $\widehat{\theta}$ to be a good estimator for θ^{causal}, which is often the most interesting estimand, we need both internal and external validity, and thus both random assignment and random sampling. □

For the binary regressor example the EHW variance estimator can be written as

$$\widehat{V}^{\text{ehw}} = \frac{N_1 - 1}{N_1^2}\widehat{S}_1^2 + \frac{N_0 - 1}{N_0^2}\widehat{S}_0^2,$$

where $\widehat{S}_1^2$ is the sample counterpart of S_1^2,

$$\widehat{S}_1^2 = \frac{1}{N_1 - 1}\sum_{i=1}^{n} R_i X_i \left(Y_i - \frac{1}{N_1}\sum_{i=1}^{n} R_i X_i Y_i\right)^2,$$

and $\widehat{S}_0^2$ is defined analogously.

If we adjust the degrees of freedom, using the modification proposed in MacKinnon and White (1985), specialized to this binary regressor example, we get the modified EHW variance estimator,

$$\widetilde{V}^{\text{ehw}} = \frac{\widehat{S}_1^2}{N_1} + \frac{\widehat{S}_0^2}{N_0},$$

with expectation equal to the sampling variance in the infinite population case,

$$V^{\text{ehw}} = E\left[\widetilde{V}^{\text{ehw}}\right] = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} = V^{\text{sampling}}(N_1, N_0, \infty, \infty).$$

This variance is also the one proposed by Neyman (1923). Bootstrapping the estimator would approximately give the same variance.

Comment 5. The EHW Variance Estimator

Compare the expected value of the modified EHW variance estimator to $V^{\text{total}}(N_1, N_0, n_1, n_0)$. There are two differences. The EHW variance $V^{\text{ehw}} = V^{\text{sampling}}(N_1, N_0, \infty, \infty)$ ignores the fact that the population may be finite, and it ignores the design component of the variance. The combination of these two differences renders the EHW variance estimator conservative:

$$V^{\text{ehw}} = V^{\text{sampling}}(N_1, N_0, \infty, \infty) \geq V^{\text{total}}(N_1, N_0, n_1, n_0).$$ □

Comment 6. Can We Improve on the EHW variance Estimator?

The difference between V^{ehw} and the total variance is

$$V^{\text{ehw}} - V^{\text{total}}(N_1, N_0, n_1, n_0) = \frac{S_\theta^2}{n}.$$

There is no good estimator for S_θ^2 because we do not observe $Y_i(1)$ and $Y_i(0)$ together. Although we may be able to come up with a lower bound for S_θ^2 that is strictly positive, there is no unbiased estimator and, typically, this term is ignored in analyses of randomized experiments (see Imbens and Rubin (2015)). In Section 3 we propose a new variance estimator that exploits the presence of fixed attributes. □

3 The General Case

This section contains the main formal results in the article. The setting we consider here allows for the presence of covariates of the causal type (*e.g.*, state institution or a regulation such as the state minimum wage), which can be discrete or continuous, as well as for the presence of covariates of the fixed attribute or characteristic type (*e.g.*, an indicator whether a state is landlocked or coastal), which again can be discrete or continuous. We allow potential causes and attributes to be systematically correlated, and we allow for general misspecification of the regression function. The conceptual difference between the causal variables and the attributes is that the value of the causal variables may depend on the design while the value of the attributes does not. This requires us to postulate potential outcomes corresponding to causes but not for the attributes.

3.1 Set Up

Consider a sequence of finite populations indexed by population size, n. Unit i in population n is characterized by a set of fixed attributes $Z_{n,i}$ (including an intercept) and by a potential outcome function, $Y_{n,i}()$, which maps causes, $U_{n,i}$, into outcomes, $Y_{n,i} = Y_{n,i}(U_{n,i})$. $Z_{n,i}$ and $U_{n,i}$ are real-valued column vectors, while $Y_{n,i}$ is scalar. We do not place restrictions on the types of the variables: they can be continuous, discrete, or mixed.

There is a sequence of samples associated with the population sequence. We will use $R_{n,i} = 1$ to indicate that unit i of population n is sampled, and $R_{n,i} = 0$ to indicate that it is not sampled. For each unit in sample n, we observe the triple, $(Y_{n,i}, U_{n,i}, Z_{n,i})$.

A key feature of the analysis in this section relative to Section 2 is that we now allow for more complicated assignment mechanisms. In particular, we relax the assumption that the causes have identical distributions.

Assumption 3. (Assignment Mechanism) *The assignments $U_{n,1}, \ldots, U_{n,n}$ are jointly independent, and independent of $R_{n,1}, \ldots, R_{n,n}$, but not (necessarily) identically distributed (i.n.i.d.).*

For what follows, it is convenient to work with a transformation $X_{n,1}, \ldots, X_{n,n}$ of $U_{n,1}, \ldots, U_{n,n}$ such that

$$E\left[\sum_{i=1}^{n} X_{n,i} Z_{n,i}'\right] = \sum_{i=1}^{n} E[X_{n,i}] Z_{n,i}' = 0. \tag{3.1}$$

This can be accomplished in the following way. We assume that the population matrix $\sum_{i=1}^{n} Z_{n,i} Z_{n,i}'$ is full-rank. Then, equation (3.1) holds for

$$X_{n,i} = U_{n,i} - \Lambda_n Z_{n,i} \tag{3.2}$$

where

$$\Lambda_n = \left(\sum_{i=1}^{n} E[U_{n,i}] Z_{n,i}'\right) \left(\sum_{i=1}^{n} Z_{n,i} Z_{n,i}'\right)^{-1}.$$

It is important to notice that, because $\Lambda_n Z_{n,i}$ is deterministic in our setting and $U_{n,1}, \ldots, U_{n,n}$ are i.n.i.d., the variables $X_{n,1}, \ldots, X_{n,n}$ are i.n.i.d. too.

For population n, let $\boldsymbol{Y}_n$, $\boldsymbol{X}_n$, $\boldsymbol{Z}_n$, $\boldsymbol{R}_n$, and $\boldsymbol{Y}_n()$ be matrices that collect outcomes, causes, attributes, sampling indicators, and potential outcome functions, where each population unit has the same row index in each of the matrices. In our setting, the sampling indicators $\boldsymbol{R}_n$ and the causes $\boldsymbol{X}_n$ are stochastic. The attributes $\boldsymbol{Z}_n$ and the potential outcome functions $\boldsymbol{Y}_n()$ are taken as fixed. Expectations are taken over the distribution of $(\boldsymbol{R}_n, \boldsymbol{X}_n)$.

We analyze the properties of the estimator $\widehat{\theta}_n$ obtained by minimizing least square errors in the sample:

$$(\widehat{\theta}_n, \widehat{\gamma}_n) = \underset{(\theta,\gamma)}{\operatorname{argmin}} \sum_{i=1}^{n} R_{n,i} \big(Y_{n,i} - X_{n,i}'\theta - Z_{n,i}'\gamma\big)^2.$$

The properties of the population regression residuals, $e_{n,i} = Y_{n,i} - X_{n,i}'\theta_n - Z_{n,i}'\gamma_n$, depend on the exact nature of the estimands, (θ_n, γ_n). In what follows, we will consider alternative target parameters, which in turn will imply different properties for $e_{n,i}$. Notice also that, although the transformation in (3.2) is typically unfeasible (because the values of $E[U_{n,i}]$ may not be known), $\widehat{\theta}_n$ is not affected by the transformation in the sense that the least squares estimators $(\widetilde{\theta}_n, \widetilde{\gamma}_n)$, defined as

$$(\widetilde{\theta}_n, \widetilde{\gamma}_n) = \underset{(\theta,\gamma)}{\operatorname{argmin}} \sum_{i=1}^{n} R_{n,i} \big(Y_{n,i} - U_{n,i}'\theta - Z_{n,i}'\gamma\big)^2,$$

satisfy $\widehat{\theta}_n = \widetilde{\theta}_n$ (although, in general, $\widehat{\gamma}_n \neq \widetilde{\gamma}_n$). As a result, we can analyze the properties of $\widehat{\theta}_n$ focusing on the properties of the regression on $X_{n,1}, \ldots, X_{n,n}$ instead of on $U_{n,1}, \ldots, U_{n,n}$.

We assume random sampling with some conditions on the sampling rate to ensure that the sample size increases with the population size.

Assumption 4. (RANDOM SAMPLING) *(i) There is a sequence of sampling probabilities, ρ_n, such that*

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

for all n-vectors $\boldsymbol{r}$ with i-th element $r_i \in \{0,1\}$. (ii) The sequence of sampling rates, ρ_n, satisfies $n\rho_n \to \infty$ and $\rho_n \to \rho \in [0,1]$.

Assumption $4(i)$ states that each population unit is sample with probability ρ_n independently of the others. The first part of Assumption $4(ii)$ guarantees that as the population size increases, the (expected) sample size also increases. The second part of Assumption $4(ii)$ allows for the possibility that asymptotically the sample size is a negligible fraction of the population size so that the EHW results, corresponding to $\rho = 0$, are included as a special case of our results.

Next assumption is a regularity condition bounding moments.

Assumption 5. (MOMENTS) *There exists some $\delta > 0$ such that the sequences*

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

are uniformly bounded.

Let

$$W_n = \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} Y_{n,i} \\ X_{n,i} \\ Z_{n,i} \end{pmatrix} \begin{pmatrix} Y_{n,i} \\ X_{n,i} \\ Z_{n,i} \end{pmatrix}', \quad \Omega_n = \frac{1}{n}\sum_{i=1}^n E\left[\begin{pmatrix} Y_{n,i} \\ X_{n,i} \\ Z_{n,i} \end{pmatrix} \begin{pmatrix} Y_{n,i} \\ X_{n,i} \\ Z_{n,i} \end{pmatrix}'\right].$$

So $\Omega_n = E[W_n]$, where the expectation is taken over the distribution of $\boldsymbol{X}_n$. We will consider also sample counterparts of W_n and Ω_n:

$$\widetilde{W}_n = \frac{1}{N}\sum_{i=1}^N R_{n,i} \begin{pmatrix} Y_{n,i} \\ X_{n,i} \\ Z_{n,i} \end{pmatrix} \begin{pmatrix} Y_{n,i} \\ X_{n,i} \\ Z_{n,i} \end{pmatrix}', \quad \widetilde{\Omega}_n = \frac{1}{N}\sum_{i=1}^n R_{n,i} E\left[\begin{pmatrix} Y_{n,i} \\ X_{n,i} \\ Z_{n,i} \end{pmatrix} \begin{pmatrix} Y_{n,i} \\ X_{n,i} \\ Z_{n,i} \end{pmatrix}'\right],$$

where $\widetilde{\Omega}_n = E[\widetilde{W}_n|\boldsymbol{R}_n]$. We will use superscripts to indicate submatrices. For example,

$$W_n = \begin{pmatrix} W_n^{YY} & W_n^{YX} & W_n^{YZ} \\ W_n^{XY} & W_n^{XX} & W_n^{XZ} \\ W_n^{ZY} & W_n^{ZX} & W_n^{ZZ} \end{pmatrix},$$

with analogous partitions for Ω_n, $\widetilde{W}_n$, and $\widetilde{\Omega}_n$. Notice that the transformation in (3.2) implies that Ω_n^{XZ} and Ω_n^{ZX} are matrices will all zero entries.

We first obtain converge results for the sample objects, $\widetilde{W}_n$ and $\widetilde{\Omega}_n$.

Lemma 1. *Suppose Assumptions 3-5 hold. Then,* $\widetilde{W}_n - \Omega_n \xrightarrow{p} 0$, $\widetilde{\Omega}_n - \Omega_n \xrightarrow{p} 0$ *and* $\widetilde{W}_n - W_n \xrightarrow{p} 0$.

Next assumption imposes convergence of second moments in the population.

Assumption 6. (Convergence of moments) $\Omega_n \to \Omega$, *which is full rank.*

3.2 Descriptive and Causal Estimands

We now define the descriptive and causal estimands that generalize θ^{descr}, $\theta^{\text{causal,sample}}$, and θ^{causal} from Section 2 to a regression context.

Definition 1. Causal and Descriptive Estimands
For a given population n, with potential outcome functions $\boldsymbol{Y}_n()$, *causes* $\boldsymbol{X}_n$, *attributes* $\boldsymbol{Z}_n$, *and sampling indicators* $\boldsymbol{R}_n$*:*

(i) Estimands are functionals of $(\boldsymbol{Y}_n(), \boldsymbol{X}_n, \boldsymbol{Z}_n, \boldsymbol{R}_n)$, *exchangeable in the rows of the arguments.*

(ii) Descriptive estimands are estimands that can be written in terms of $\boldsymbol{Y}_n$, $\boldsymbol{X}_n$, *and* $\boldsymbol{Z}_n$, *free of dependence on* $\boldsymbol{R}_n$, *and free of dependence on* $\boldsymbol{Y}_n()$ *beyond dependence on* $\boldsymbol{Y}_n$.

(iii) Causal estimands are estimands that cannot be written in terms of $\boldsymbol{Y}_n$, $\boldsymbol{X}_n$, $\boldsymbol{Z}_n$, *and* $\boldsymbol{R}_n$, *because they depend on the potential outcome functions* $\boldsymbol{Y}_n()$ *beyond the realized outcomes,* $\boldsymbol{Y}_n$.

Causal estimands depend on the values of potential outcomes beyond the values that can be inferred from the realized outcomes. Given a sample, the only reason we may not be able to infer the value of a descriptive estimand is that we may not see all the units in the population. In contrast, even if we observe all units in a population, we may not be able to infer the value of a causal estimand because its value depends on potential outcomes.

We define three estimands of interest,

$$\begin{pmatrix} \theta_n^{\text{descr}} \\ \gamma_n^{\text{descr}} \end{pmatrix} = \begin{pmatrix} W_n^{XX} & W_n^{XZ} \\ W_n^{ZX} & W_n^{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} W_n^{XY} \\ W_n^{ZY} \end{pmatrix}, \tag{3.3}$$

$$\begin{pmatrix} \theta_n^{\text{causal,sample}} \\ \gamma_n^{\text{causal,sample}} \end{pmatrix} = \begin{pmatrix} \widetilde{\Omega}_n^{XX} & \widetilde{\Omega}_n^{XZ} \\ \widetilde{\Omega}_n^{ZX} & \widetilde{\Omega}_n^{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} \widetilde{\Omega}_n^{XY} \\ \widetilde{\Omega}_n^{ZY} \end{pmatrix}, \tag{3.4}$$

and

$$\begin{pmatrix} \theta_n^{\text{causal}} \\ \gamma_n^{\text{causal}} \end{pmatrix} = \begin{pmatrix} \Omega_n^{XX} & \Omega_n^{XZ} \\ \Omega_n^{ZX} & \Omega_n^{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} \Omega_n^{XY} \\ \Omega_n^{ZY} \end{pmatrix}. \tag{3.5}$$

Alternatively, the estimands in (3.3) to (3.5) can be defined as the coefficients that correspond to the orthogonality conditions in terms of the residuals $e_{n,i} = Y_{n,i} - X_{n,i}'\theta_n - Z_{n,i}'\gamma_n$,

$$\frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} X_{n,i} \\ Z_{n,i} \end{pmatrix} e_{n,i} = 0, \quad \frac{1}{n}\sum_{i=1}^{n} R_{n,i} E\left[\begin{pmatrix} X_{n,i} \\ Z_{n,i} \end{pmatrix} e_{n,i} \right] = 0, \quad \frac{1}{n}\sum_{i=1}^{n} E\left[\begin{pmatrix} X_{n,i} \\ Z_{n,i} \end{pmatrix} e_{n,i} \right] = 0,$$

respectively. We will study the properties of the least squares estimator, $\widehat{\theta}_n$, defined by

$$\begin{pmatrix} \widehat{\theta}_n \\ \widehat{\gamma}_n \end{pmatrix} = \begin{pmatrix} \widetilde{W}_n^{XX} & \widetilde{W}_n^{XZ} \\ \widetilde{W}_n^{ZX} & \widetilde{W}_n^{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} \widetilde{W}_n^{XY} \\ \widetilde{W}_n^{ZY} \end{pmatrix},$$

as an estimator of the parameters defined in equations (3.3) to (3.5).

Notice that, by the law of total expectation, and because the potential outcome functions are fixed in our framework, $\theta_n^{\text{causal,sample}}$ and θ_n^{causal} are causal estimands, according to our definition, while, θ_n^{descr} is not. The fact that an estimand is causal according to our definition does not imply it has an interpretation as an average causal effect. In Section 3.3 we present conditions under which the regression estimand does have such an interpretation.

3.3 Causal Interpretations of the Estimands

By construction, the descriptive estimand can be interpreted as the set of coefficients of a population best linear predictor (least squares). A more challenging question concerns the interpretation of the two causal estimands, and in particular their relation to the potential outcome functions. In this section we investigate this question.

The first part of our proposed set of conditions for a causal interpretation of $\theta_n^{\text{causal,sample}}$ and θ_n^{causal} generalizes random assignment.

Assumption 7. (Linearity of the Expected Assignment) *There exists a sequence of real matrices B_n such that*

$$E[U_{n,i}] = B_n Z_{n,i}.$$

for n large enough.

Because of Lemma 1 and Assumption 6, $\Lambda_n = B_n$, which implies that $E[X_{n,i}] = 0$. Assumption 7 looks very different from conventional exogeneity or unconfoundedness conditions, where the residuals, $e_{n,i}$, are assumed to be (mean-) independent of the regressors, and so it merits some discussion. A special case of this assumption arises in the context of a randomized assignment, when $E[U_{n,i}]$ is constant across units. In that case Assumption 7 holds as long as there is an intercept in the set of attributes. More generally, Assumption 7 relaxes the completely randomized assignment setting, by allowing the distribution of $U_{n,i}$ to depend on the attributes. However, this dependence is restricted in that the mean of $U_{n,i}$ is linear in $Z_{n,i}$. For example, Assumption 7 holds automatically when $U_{n,1}, \ldots, U_{n,n}$ are identically distributed and $Z_{n,i}$ contains a saturated set on indicators for all possible values of the attributes. Later in this section, we will show that under a set of conditions that includes Assumption 7, θ_n^{causal} and $\theta^{\text{causal,sample}}$ can be interpreted as weighted averages of unit-level causal effects. The connection between linearity in the "propensity score" (Rosenbaum and Rubin (1983), in our analysis represented by $E[U_{n,i}] = B_n Z_{n,i}$) and the interpretation of population regression coefficients as weighted averages of heterogeneous causal effects has been previously noticed in related contexts (see Angrist, 1998; Angrist and Pischke, 2008; Aronow and Samii, 2016; Słoczyński, 2017).

Assumption 8. (Linearity of Potential Outcomes)

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

almost surely, where $\theta_{n,i}$ and $\xi_{n,i}$ are non-stochastic.

In this formulation, any dependence of $Y_{n,i}$ on $Z_{n,i}$ or on unobserved attributes is subsumed by $\theta_{n,i}$ and $\xi_{n,i}$, which are non-stochastic. Each element of the vector $\theta_{n,i}$ represents the causal effect on $Y_{n,i}$ of increasing the corresponding value of $U_{n,i}$ in one unit.

Theorem 1. *Suppose Assumptions 3-8 hold. Then, for n large enough,*

$$\theta_n^{\text{causal}} = \left(\sum_{i=1}^{n} E\left[W_{n,i}^{XX}\right] \right)^{-1} \sum_{i=1}^{n} E\left[W_{n,i}^{XX}\right] \theta_{n,i},$$

and, with probability approaching one,

$$\theta_n^{\text{causal,sample}} = \left(\sum_{i=1}^{n} R_{n,i} E\left[W_{n,i}^{XX} \right] \right)^{-1} \sum_{i=1}^{n} R_{n,i} E\left[W_{n,i}^{XX} \right] \theta_{n,i},$$

where $W_{n,i}^{XX} = X_{n,i} X_{n,i}'$.

The linearity in Assumption 8 is a strong restriction in many settings. In some other settings, in particular, when the causal variable is binary or, more generally when the causal variable takes on only a finite number of values, it is immediate to enforce this assumption by including in $U_{n,i}$ indicator variables representing each but one of the possible values of the cause. Assumption 8 can be relaxed at the cost of introducing additional complication in the interpretation of the estimands.

Theorem 2. *Suppose that Assumptions 3-7 hold. Moreover, assume that* $X_{n,1}, \ldots, X_{n,n}$ *are continuous random variables with convex and compact supports, and that the potential outcome functions,* $Y_{n,i}()$ *are continuously differentiable. Then, there exist random variables* $v_{n,1}, \ldots, v_{n,n}$ *such that, for* n *sufficiently large,*

$$\theta_n^{\text{causal}} = \left(\sum_{i=1}^{n} E\left[W_{n,i}^{XX} \right] \right)^{-1} \sum_{i=1}^{n} E\left[W_{n,i}^{XX} \varphi_{n,i} \right],$$

and

$$\theta_n^{\text{causal,sample}} = \left(\sum_{i=1}^{n} R_{n,i} E\left[W_{n,i}^{XX} \right] \right)^{-1} \sum_{i=1}^{n} R_{n,i} E\left[W_{n,i}^{XX} \varphi_{n,i} \right],$$

where $\varphi_{n,i}$ *is the derivative of* $Y_{n,i}()$ *evaluated at* $v_{n,i}$.

Comment 7. Here, we provide a simple example that shows how the result in Theorems 1 and 2 may not hold in the absence of Assumption 7. Consider the population with three units described

Table 3

Unit	$Y_i(x)$	Z_i	$E[U_i]$	$\text{var}(U_i)$
1	a	-1	b	1
2	0	0	$-2b$	1
3	$2a$	1	b	1

in Table 3 (where, for simplicity, we drop the subscript n). In this example, $E[U_i] = 3bZ_i^2 - 2b$

is a non-linear function of Z_i. Notice that

$$\sum_{i=1}^{3} E[U_i]/3 = \sum_{i=1}^{3} E[U_i]Z_i/3 = 0,$$

so that $X_i = U_i$. Therefore, $E[X_i^2] = E[U_i^2]$. Also, because potential outcomes do not depend on X_i, it follows that $E[X_iY_i] = E[X_i]Y_i = E[U_i]Y_i$. As a result,

$$\theta^{\text{causal}} = \left(\sum_{i=1}^{3} E[X_i^2]\right)^{-1} \sum_{i=1}^{3} E[X_iY_i] = \frac{ab}{2b^2+1},$$

which is different from zero as long as $ab \neq 0$. In this example all the potential outcome functions $Y_i(x)$ are flat as a function of x, so all unit-level causal effects of the type $Y_i(x) - Y_i(x')$ are zero, and yet the causal least squares estimand can be positive or negative depending on the values of a and b. □

3.4 The Asymptotic Distribution of The Least Squares Estimator

In this section we present one of the main results of the article, describing the properties of the least squares estimator viewed as an estimator of the causal estimands and, separately, viewed as an estimator of the descriptive estimand. In contrast to Section 2, we do not have exact results, relying instead on asymptotic results based on sequences of populations.

First, we define the population residuals, denoted by $\varepsilon_{n,i}$, relative to the population causal estimands,

$$\varepsilon_{n,i} = Y_{n,i} - X'_{n,i}\theta_n^{\text{causal}} - Z'_{n,i}\gamma_n^{\text{causal}}. \tag{3.6}$$

Comment 8. The definition of the residuals, $\varepsilon_{n,1}, \ldots, \varepsilon_{n,n}$, mirrors that in conventional regression analysis, but their properties are conceptually different. For instance, the residuals need not be stochastic. If they are stochastic, they are so because of their dependence on $\boldsymbol{X}_n$. □

Under the assumption that the $X_{n,i}$ are jointly independent (but not necessarily identically distributed), the n products $X_{n,i}\varepsilon_{n,i}$ are jointly independent but not identically distributed. Most importantly, in general the expectations $E[X_{n,i}\varepsilon_{n,i}]$ may vary across i, and need not all be zero. However, as shown in Section 3.2, the *averages* of these expectations over the entire population are guaranteed to be zero by the definition of $(\theta_n^{\text{causal}}, \gamma_n^{\text{causal}})$. Define the limits of

the population variance,

$$\Delta^{\text{cond}} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \text{var}\left(X_{n,i}\varepsilon_{n,i}\right),$$

and the expected outer product

$$\Delta^{\text{ehw}} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E\left[X_{n,i}\varepsilon_{n,i}^2 X_{n,i}'\right].$$

The difference between Δ^{ehw} and Δ^{cond} is the limit of the average outer product of the means,

$$\Delta^{\mu} = \Delta^{\text{ehw}} - \Delta^{\text{cond}} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E[X_{n,i}\varepsilon_{n,i}]E[X_{n,i}\varepsilon_{n,i}]',$$

which is positive semidefinite. We assume existence of these limits.

Assumption 9. (Existence of Limits) Δ^{cond} *and* Δ^{ehw} *exist and are positive definite.*

Theorem 3. *Suppose Assumptions 3-9 hold, and let* $H = \Omega^{XX} = \lim_{n\to\infty} \Omega_n^{XX}$. *Then,*

(i)

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

(ii)

$$\sqrt{N}\left(\widehat{\theta}_n - \theta_n^{\text{causal,sample}}\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1}\Delta^{\text{cond}}H^{-1}\right),$$

(iii)

$$\sqrt{N}\left(\widehat{\theta}_n - \theta_n^{\text{descr}}\right) \xrightarrow{d} \mathcal{N}\left(0, (1-\rho)H^{-1}\Delta^{\text{ehw}}H^{-1}\right).$$

Comment 9. For both the population causal and the descriptive estimand the asymptotic variance in the case with $\rho = 0$ reduces to the standard EHW variance, $H^{-1}\Delta^{\text{ehw}}H^{-1}$. If the sample size is non-negligible as a fraction of the population size, $\rho > 0$, the difference between the EHW variance and the finite population causal variance is positive semi-definite and equal to $\rho H^{-1}(\Delta^{\text{ehw}} - \Delta^{\text{cond}})H^{-1}$. □

3.5 The Variance Under Correct Specification

Consider a constant treatment effect assumption, which is required for a correct specification of a linear regression function as a function that describes potential outcomes.

Assumption 10. (Constant Treatment Effects)

$$Y_{n,i} = U_{n,i}'\theta_n + \xi_{n,i}$$

almost surely, where θ_n *and* $\xi_{n,i}$ *are non-stochastic.*

This strengthens Assumption 8 by requiring that the $\theta_{n,i}$ do not vary by i.

Under Assumption 10, Theorem 1 implies that $\theta_n^{\text{causal}} = \theta_n$ (although it need not be the case that $\theta^{\text{descr}} = \theta_n$). Then, for

$$\lambda_n = \left(\sum_{i=1}^{n} Z_{n,i}Z_{n,i}'\right)^{-1} \sum_{i=1}^{n} Z_{n,i}\xi_{n,i}$$

we obtain that equation (3.6) holds for $\gamma_n^{\text{causal}} = \Lambda_n'\theta_n + \lambda_n$ and $\varepsilon_{n,i} = \xi_{n,i} - Z_{n,i}'\lambda_n$. In this case, the residuals, $\varepsilon_{n,i}$, are non-stochastic. As a result, $E[X_{n,i}\varepsilon_{n,i}] = E[X_{n,i}]\varepsilon_{n,i} = 0$, with implies $\Delta^{\mu} = \Delta^{\text{ehw}} - \Delta^{\text{cond}} = 0$. This leads to the following result.

Theorem 4. *Suppose that Assumptions 3-10 hold. Then,*

$$\sqrt{N}\left(\widehat{\theta} - \theta_n^{\text{causal}}\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1}\Delta^{\text{ehw}}H^{-1}\right),$$

irrespective of the value of ρ.

Notice that the result of the theorem applies also with $\theta_n^{\text{causal,sample}}$ replacing θ_n^{causal} because the two parameter vectors are identical (with probability approaching one) under Assumption 10.

Comment 10. The key insight in this theorem is that the asymptotic variance of $\widehat{\theta}_n$ does not depend on the ratio of the sample to the population size when the regression function is correctly specified. Therefore, it follows that the usual EHW variance matrix is correct for $\widehat{\theta}_n$ under these assumptions. For the case with $X_{n,i}$ binary and no attributes beyond the intercept, this result can be inferred directly from Neyman's results for randomized experiments (Neyman, 1923). In that case, the result of Theorem 4 follows from the restriction of constant treatment effects,

$Y_{n,i}(1) - Y_{n,i}(0) = \theta_n$, which is extended to the more general case of non-binary regressors in Assumption 10. The asymptotic variance of $\widehat{\gamma}_n$, the least squares estimator of the coefficients on the attributes, still depends on the ratio of sample to population size, and it can be shown that the conventional robust EHW estimator continues to over-estimate the variance of $\widehat{\gamma}_n$. □

4 Estimating the Variance

Now let us turn to the problem of estimating the variance for the descriptive and causal estimands. There are four components to the asymptotic variance, ρ, H, Δ^{ehw} and Δ^{cond}. The first three are straightforward to estimate. The ratio ρ can be estimated as N/n. To estimate H, first estimate Λ_n as

$$\widehat{\Lambda}_n = \left(\sum_{i=1}^n R_{n,i} U_{n,i} Z_{n,i}'\right) \left(\sum_{i=1}^n R_{n,i} Z_{n,i} Z_{n,i}'\right)^{-1}.$$

Then one can estimate H as the average of the matrix of outer products over the sample:

$$\widehat{H}_n = \frac{1}{N} \sum_{i=1}^n R_{n,i} \left(U_{n,i} - \widehat{\Lambda}_n Z_{n,i}\right) \left(U_{n,i} - \widehat{\Lambda}_n Z_{n,i}\right)'.$$

It is also straightforward to estimate Δ^{ehw}. First we estimate the residuals for the units in the sample, $\widehat{\varepsilon}_{n,i} = Y_{n,i} - (U_{n,i} - \widehat{\Lambda}_n Z_{n,i})'\widehat{\theta}_n - Z_{n,i}'\widehat{\gamma}_n$, and then we estimate Δ^{ehw} as:

$$\widehat{\Delta}_n^{\text{ehw}} = \frac{1}{N} \sum_{i=1}^n R_{n,i} (U_{n,i} - \widehat{\Lambda}_n Z_{n,i})\, \widehat{\varepsilon}_{n,i}^2\, (U_{n,i} - \widehat{\Lambda}_n Z_{n,i})'.$$

The EHW variance, $V^{\text{ehw}} = H^{-1}\Delta^{\text{ehw}}H^{-1}$, is then estimated as

$$\widehat{V}_n^{\text{ehw}} = \widehat{H}_n^{-1} \widehat{\Delta}_n^{\text{ehw}} \widehat{H}_n^{-1}.$$

Lemma 2. *Suppose Assumptions 3-7 and 9 hold with $\delta = 4$. Then,*

$$\widehat{V}_n^{\text{ehw}} \xrightarrow{p} V^{\text{ehw}}.$$

Alternatively one can use resampling methods such as the bootstrap (e.g., Efron, 1987).

It is more challenging to estimate Δ^{cond}. The reason is the same that makes it impossible to obtain unbiased estimates of the variance of the estimator for the average treatment effect in the example in Section 2. In that case there are three terms in the expression for the variance in

equation (2.4). The first two are straightforward to estimate, but the third one, S_θ^2/n cannot be estimated consistently because we do not observe both potential outcomes for the same units. Often, researchers use the conservative estimator based on ignoring S_θ^2/n. If we proceed in the same fashion for the regression context of Section 3, we obtain the conservative estimator $\widehat{V}^{\text{ehw}}$, based on ignoring Δ^μ. We show, however, that in the presence of attributes we can improve the variance estimator. We build on Abadie and Imbens (2008), Abadie et al. (2014), and Fogarty (2016) who, in contexts different than the one studied in this article, have used the explanatory power of attributes to improve variance estimators. While Abadie and Imbens (2008), Abadie et al. (2014) use nearest-neighbor techniques, here we follow Fogarty (2016) and apply linear regression techniques. The proposed estimator replaces the expectations $E[X_{n,i}\varepsilon_{n,i}]$, which cannot be consistently estimated, with predictors from a linear least squares projection of estimates of $X_{n,i}\varepsilon_{n,i}$ on the attributes, $Z_{n,i}$. Let $\widehat{X}_{n,i} = U_{n,i} - \widehat{\Lambda}_n Z_{n,i}$, and

$$\widehat{G}_n = \left(\frac{1}{N}\sum_{i=1}^{n} R_{n,i}\widehat{X}_{n,i}\widehat{\varepsilon}_{n,i}Z'_{n,i}\right)\left(\frac{1}{N}\sum_{i=1}^{n} R_{n,i}Z_{n,i}Z'_{n,i}\right)^{-1}.$$

The matrix $\widehat{G}_n$ contains the coefficients of a least squares regression of $\widehat{X}_{n,i}\widehat{\varepsilon}_{n,i}$ on $Z_{n,i}$. The next assumption ensures convergence of $\widehat{G}_n$.

Assumption 11.

$$\frac{1}{n}\sum_{i=1}^{n} E[X_{n,i}\varepsilon_{n,i}]Z'_{n,i}$$

has a limit.

Consider now the following estimator,

$$\widehat{\Delta}_n^Z = \frac{1}{N}\sum_{i=1}^{n} R_{n,i}\left(\widehat{X}_{n,i}\widehat{\varepsilon}_{n,i} - \widehat{G}_n Z_{n,i}\right)\left(\widehat{X}_{n,i}\widehat{\varepsilon}_{n,i} - \widehat{G}_n Z_{n,i}\right)'.$$

which uses $\widehat{G}_n Z_{n,i}$ in lieu of a consistent estimator of $E[X_{n,i}\varepsilon_{n,i}]$. Notice that we do not assume that $E[X_{n,i}\varepsilon_{n,i}]$ is linear in $Z_{n,i}$. However, we will show that, as long as the attributes can linearly explain some of the variance in $\widehat{X}_{n,i}\widehat{\varepsilon}_{n,i}$, the estimator $\widehat{\Delta}_n^Z$ is smaller (in a matrix sense) than $\widehat{\Delta}_n^{\text{ehw}}$. Moreover, $\widehat{\Delta}_n^Z$ remains conservative in large samples. These results are provided in the following lemma.

Lemma 3. *Suppose Assumptions 3-7, 9 and 11 hold with $\delta = 4$. Then, $0 \leq \widehat{\Delta}_n^Z \leq \widehat{\Delta}_n^{\text{ehw}}$, and $\widehat{\Delta}_n^Z \xrightarrow{p} \Delta^Z$, where $\Delta^{\text{cond}} \leq \Delta^Z \leq \Delta^{\text{ehw}}$ (all inequalities are to be understood in a matrix sense).*

Variance estimators follow immediately from Lemma 3 by replacing Δ^{cond} with the estimate $\widehat{\Delta}_n^Z$ in the asymptotic variance formulas of Theorem 3. These estimators are not larger (and typically smaller) than estimators based on $\widehat{\Delta}_n^{\text{ehw}}$, and they remain conservative in large samples. For simplicity, Lemma 3 is based on a linear predictor for $E[X_{n,i}\varepsilon_{n,i}]$. Modifications that accommodate nonlinear predictors are immediate, at the cost of additional assumptions.

Comment 11. A special case of the adjusted variance estimate is an estimate obtained from stratifying the sample on the basis of attributes $Z_{n,i}$. In particular, if $Z_{n,i}$ includes exhaustive, mutually exclusive dummy variables – or, if we reduce the information in $Z_{n,i}$ down to such indicators – then $\hat{\Delta}_n^Z$ reduces to the middle of the sandwich in a commonly used estimator in the context of standard stratified sampling. (See, for example, (Wooldridge (2010), Section 20.2.2).) Then, the residuals from regressing $\hat{X}_{n,i}\hat{\varepsilon}_{n,i}$ on $Z_{n,i}$ are simply stratum-specific demeaned versions of $\hat{X}_{n,i}\hat{\varepsilon}_{n,i}$. Such a variance estimator is easy to obtain using standard software packages that support regression with survey samples. □

5 Inference for Alternative Questions

This article has focused on inference for descriptive and causal estimands in a single cross-section. For example, we might have a sample that includes outcomes from all countries in a particular year, say 2013. In words, we analyze inference for estimands of parameters that answer the following causal question: "What is the difference between what the average outcome would have been in those countries in the year 2013 if all had been treated, and what the average outcome would have been if all had not been treated?" We also analyze inference for estimands of parameters that can be used to answer descriptive questions, such as "What was the difference in outcomes between Northern and Southern countries in the year 2013?"

These are not the only questions a researcher could focus on. An alternative question might be, "what is the expected difference in average outcomes between Northern and Southern countries in a future year, say the year 2020," or "what is the difference between what the average outcome would be in those countries in the year 2020 if all would be treated, and what the average outcome would be if none would be treated?" Arguably in most empirical analyses that are

intended to inform policy the object of interest depends on future, not simply on past, outcomes. This creates substantial problems for inference. Here we discuss some of the complications, but much of this is left for future work. Our two main points are, first, that it is important to be explicit about the estimand, and second, that the conventional robust standard errors were not designed to solve these problems and do not do so without strong, typically implausible, assumptions.

Formally questions that involve future values of outcomes for countries could be formulated in terms of a population of interest that includes as its units each country in a variety of different states of the world that might be realized in future years. This population is large if there are many possible realizations of states of the world (e.g., rainfall, local political conditions, natural resource discoveries, etc.), with a potentially complex dependence structure. Given such a population the researcher may wish to estimate, say the difference in average 2020 outcomes for two sets of countries, and calculate standard errors based on values for the outcomes for the same set of countries in an earlier year, say 2020. A natural estimator for the difference in average values for Northern and Southern countries in 2020 would be the corresponding difference in average values in 2013. However, even though such data would allow us to infer without uncertainty the difference in average outcomes for Northern and Southern countries in 2013, there would be uncertainty regarding the true value of that difference in the year 2020. In order to construct confidence intervals for the difference in 2020, the researcher must make some assumptions about how country outcomes will vary from year to year. An extreme assumption is that outcomes in 2013 and 2020 for the same country are independent conditional on attributes, which would justify the conventional EHW variance estimator. However, assuming that there is no correlation between outcomes for the same country in successive years appears highly implausible. In fact any assumption about the magnitude of this correlation in the absence of direct information about it in the form of panel data would appear to be controversial. Such assumptions would also depend heavily on the future year for which we would wish to estimate the difference in averages, again highlighting the importance of being precise about the estimand.

Although in this case there is uncertainty regarding the difference in average outcomes in 2020 despite the fact that the researchers observes (some) information on all countries in the population of interest, we emphasize that the assumptions required to validate the application of EHW standard errors in this setting are strong and arguably implausible. Moreover, researchers rarely formally state the population of interest, let alone state and justify the assumptions that

justify inference.

Generally, if future predictions are truly the primary question of interest, it seems prudent to explicitly state the assumptions that justify particular calculations for standard errors. Especially in the absence of panel data, the results are likely to be sensitive to such assumptions. With panel data the researcher may be able to estimate the dynamic process underlying the potential outcomes in order to obtain standard errors for the future predictions. In practice it may be useful to report standard errors for various estimands. For example, if the primary estimand is an average causal effect in the future, it may still be useful to report estimates and standard errors for the same contemporaneous average causal effect, in combination with estimates and standard errors for the future average causal effect, in order to understand the additional uncertainty that comes with predictions for a future period. We leave this direction for future work.

6 Conclusion

In this article we study the interpretation of standard errors in regression analysis when the assumption that the sample is drawn randomly from a large population of interest is not attractive. The conventional robust standard errors justified by the random sampling assumption do not necessarily apply in this case. We show that, by viewing covariates as potential causes in a Rubin Causal Model or potential outcome framework, we can provide a coherent interpretation for standard errors that allows for uncertainty coming from both random sampling and from conditional random assignment. The proposed standard errors may be different from the conventional ones.

In the current article we focus exclusively on regression models, and we provide a full analysis of inference for only a certain class of regression models with some of the covariates causal and some attributes. Thus, this article is only a first step in a broader research program. The concerns we have raised in this article arise in many other settings and for other kinds of hypotheses, and the implications would need to be worked out for those settings. Section 5 suggests some directions we think are particularly natural to consider.

Appendix

I. A Bayesian Approach

Given that we are advocating for a different conceptual approach to modeling inference, it is useful to look at the problem from more than one perspective. In this section we consider a Bayesian perspective and re-analyze the example from Section 2. Using a simple parametric model we show that in a Bayesian approach the same issues arise in the choice of estimand. Viewing the problem from a Bayesian perspective reinforces the point that formally modeling the population and the sampling process leads to the conclusion that inference is different for descriptive and causal questions. Note that in this discussion the notation will necessarily be slightly different from the rest of the article; notation and assumptions introduced in this subsection apply only within this subsection.

Define $\boldsymbol{Y}_n(1)$, $\boldsymbol{Y}_n(0)$ to be the n vectors with typical elements $Y_i(1)$ and $Y_i(0)$, respectively. We view the n-vectors $\boldsymbol{Y}_n(1)$, $\boldsymbol{Y}_n(0)$, $\boldsymbol{R}_n$, and $\boldsymbol{X}_n$ as random variables, some observed and some unobserved. We assume the rows of the $n \times 4$ matrix $[\boldsymbol{Y}_n(1), \boldsymbol{Y}_n(0), \boldsymbol{R}_n, \boldsymbol{X}_n]$ are exchangeable. Then, by appealing to DeFinetti's theorem, we model this, with no essential loss of generality (for large n) as the product of n independent and identically distributed random quadruples $(Y_i(1), Y_i(0), R_i, X_i)$ given some unknown parameter β:

$$f(\boldsymbol{Y}_n(1), \boldsymbol{Y}_n(0), \boldsymbol{R}_n, \boldsymbol{X}_n) = \prod_{i=1}^{n} f(Y_i(1), Y_i(0), R_i, X_i|\beta).$$

Inference then proceeds by specifying a prior distribution for β, say $p(\beta)$. To make this specific, consider following model. Let X_i and R_i have Binomial distributions with parameters q and ρ,

$$\Pr(X_i = 1|Y_i(1), Y_i(0), R_i) = q, \qquad \Pr(R_i = 1|Y_i(1), Y_i(0)) = \rho.$$

The pairs $(Y_i(1), Y_i(0))$ are assumed to be jointly normally distributed:

$$\left(\begin{array}{c} Y_i(1) \\ Y_i(0) \end{array} \right) \Bigg| \, \mu_1, \mu_0, \sigma_1^2, \sigma_0^2, \kappa \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_1 \\ \mu_0 \end{array} \right), \left(\begin{array}{cc} \sigma_1^2 & \kappa\sigma_1\sigma_0 \\ \kappa\sigma_1\sigma_0 & \sigma_1^2 \end{array} \right) \right),$$

so that the full parameter vector is $\beta = (q, \rho, \mu_1, \mu_0, \sigma_1^2, \sigma_0^2, \kappa)$.

We change the observational scheme slightly from Section 2 to allow for the analytic derivation of posterior distributions. We assume that for all units in the population we observe the pair (R_i, X_i), and for units with $R_i = 1$ we observe the outcome $Y_i = Y_i(X_i)$. Define $\widetilde{Y}_i = R_i Y_i$, so for all units in the population we observe the triple $(R_i, X_i, \widetilde{Y}_i)$. Let $\boldsymbol{R}_n$, $\boldsymbol{X}_n$, and $\widetilde{\boldsymbol{Y}}_n$ be the n vectors of these variables. $\bar{Y}_1$ denotes the average of Y_i in the subpopulation with $R_i = 1$ and $X_i = 1$, and $\bar{Y}_0$ denotes the average of Y_i in the subpopulation with $R_i = 1$ and $X_i = 0$.

The descriptive estimand is

$$\theta_n^{\text{descr}} = \frac{1}{n_1} \sum_{i=1}^{n} X_i Y_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - X_i) Y_i.$$

The causal estimand is

$$\theta_n^{\text{causal}} = \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i(1) - Y_i(0) \Big).$$

It is interesting to compare these estimands to an additional estimand, the super-population average treatment effect,

$$\theta^{\text{causal}} = \mu_1 - \mu_0.$$

In general these three estimands are distinct, with their own posterior distributions, but in some cases, notably when n is large, the three posterior distributions are similar.

It is instructive to consider a very simple case where analytic solutions for the posterior distribution for θ_n^{descr}, θ_n^{causal}, and θ^{causal} are available. Suppose σ_1^2, σ_0^2, κ and q are known, so that the only unknown parameters are the two means μ_1 and μ_0. Finally, let us use independent, diffuse (improper), prior distributions for μ_1 and μ_0.

Then, a standard result is that the posterior distribution for (μ_1, μ_0) given $(\boldsymbol{R}_n, \boldsymbol{X}_n, \widetilde{\boldsymbol{Y}}_n)$ is

$$\left(\begin{array}{c} \mu_1 \\ \mu_0 \end{array} \right) \Bigg| \, \boldsymbol{R}_n, \boldsymbol{X}_n, \widetilde{\boldsymbol{Y}}_n \sim \mathcal{N}\left(\left(\begin{array}{c} \bar{Y}_1 \\ \bar{Y}_0 \end{array} \right), \left(\begin{array}{cc} \sigma_1^2/N_1 & 0 \\ 0 & \sigma_0^2/N_0 \end{array} \right) \right),$$

where N_1 is the number of units with $R_i = 1$ and $X_i = 1$, and N_0 is the number of units with $R_i = 1$ and $X_i = 0$. This directly leads to the posterior distribution for θ^{causal}:

$$\theta^{\text{causal}} | \boldsymbol{R}_n, \boldsymbol{X}_n, \widetilde{\boldsymbol{Y}}_n \sim \mathcal{N}\left(\bar{Y}_1 - \bar{Y}_0, \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0} \right).$$

A longer calculation leads to the posterior distribution for the descriptive estimand:

$$\theta_n^{\text{descr}} | \boldsymbol{R}_n, \boldsymbol{X}_n, \widetilde{\boldsymbol{Y}}_n \sim \mathcal{N}\left(\bar{Y}_1 - \bar{Y}_0, \frac{\sigma_1^2}{N_1}\left(1 - \frac{N_1}{n_1}\right) + \frac{\sigma_0^2}{N_0}\left(1 - \frac{N_0}{n_0}\right) \right).$$

The implied posterior interval for θ_n^{descr} is very similar to the corresponding confidence interval based on the normal approximation to the sampling distribution for $\bar{Y}_1 - \bar{Y}_0$. If n_1 and n_0 are large, this posterior distribution is close to the posterior distribution of the causal estimand. If, on the other hand, $N_1 = n_1$ and $N_0 = n_0$, then the posterior distribution of the descriptive estimand becomes degenerate and centered at $\bar{Y}_1 - \bar{Y}_0$.

A somewhat longer calculation for θ_n^{causal} leads to

$$\begin{aligned} \theta_n^{\text{causal}} | \boldsymbol{R}_n, \boldsymbol{X}_n, \widetilde{\boldsymbol{Y}}_n \sim \mathcal{N}\Bigg(& \bar{Y}_1 - \bar{Y}_0, \frac{N_0}{n^2}\sigma_1^2(1-\kappa^2) + \frac{N_1}{n^2}\sigma_0^2(1-\kappa^2) \\ & + \frac{n-N}{n^2}\sigma_1^2 + \frac{n-N}{n^2}\sigma_0^2 - 2\frac{n-N}{n^2}\kappa\sigma_1\sigma_0 \\ & + \frac{\sigma_1^2}{N_1}\left(1 - \left(1 - \kappa\frac{\sigma_0}{\sigma_1}\right)\frac{N_1}{n}\right)^2 + \frac{\sigma_0^2}{N_0}\left(1 - \left(1 - \kappa\frac{\sigma_1}{\sigma_0}\right)\frac{N_0}{n}\right)^2 \Bigg). \end{aligned}$$

Consider the special case of constant treatment effects, where $Y_i(1) - Y_i(0) = \mu_1 - \mu_0$. Then, $\kappa = 1$, and $\sigma_1 = \sigma_0$, and the posterior distribution of θ_n^{causal} is the same as the posterior distribution of θ^{causal}. The same posterior distribution arises in the limit if n goes to infinity, regardless of the values of κ, σ_1, and σ_0.

To sum up, if the population is large, relative to the sample, the posterior distributions of θ_n^{descr}, θ_n^{causal} and θ^{causal} agree. However, if the population is small, the three posterior distributions differ, and the

researcher needs to be precise in defining the estimand. In such cases, simply focusing on the super-population estimand $\theta^{\text{causal}} = \mu_1 - \mu_0$ is arguably not appropriate, and the posterior inferences for such estimands will differ from those for other estimands such as θ_n^{causal} or θ_n^{descr}.

II. Proofs

Proof of Lemma 1: See supplementary appendix. □

Proof of Theorem 1: Under the stated conditions, the matrices $\sum_{i=1}^n Z_{n,i}Z'_{n,i}$ and $\sum_{i=1}^n R_{n,i}Z_{n,i}Z'_{n,i}$ are invertible with probability approaching one. As a result, with probability approaching one

$$\begin{aligned} B_n &= \left(\sum_{i=1}^n R_{n,i}E[U_{n,i}]Z'_{n,i}\right)\left(\sum_{i=1}^n R_{n,i}Z_{n,i}Z'_{n,i}\right)^{-1} \\ &= \left(\sum_{i=1}^n E[U_{n,i}]Z'_{n,i}\right)\left(\sum_{i=1}^n Z_{n,i}Z'_{n,i}\right)^{-1} = \Lambda_n. \end{aligned}$$

As a result, we obtain

$$\theta_n^{\text{causal}} = \left(\sum_{i=1}^n E[X_{n,i}X'_{n,i}]\right)^{-1}\sum_{i=1}^n E[X_{n,i}Y_{n,i}],$$

and

$$\theta_n^{\text{causal,sample}} = \left(\sum_{i=1}^n R_{n,i}E[X_{n,i}X'_{n,i}]\right)^{-1}\sum_{i=1}^n R_{n,i}E[X_{n,i}Y_{n,i}].$$

Now,

$$\begin{aligned} E[X_{n,i}Y_{n,i}] &= E[X_{n,i}U'_{n,i}]\theta_{n,i} + E[X_{n,i}]\xi_{n,i} \\ &= E[X_{n,i}X'_{n,i}]\theta_{n,i}. \end{aligned}$$

implies the results. □

Proof of Theorem 2: Let $\nabla Y_{n,i}()$ be the gradient of $Y_{n,i}()$. By the mean value theorem there exists sets $\mathcal{T}_{n,i} \subseteq [0,1]$ such that for any $t_{n,i} \in \mathcal{T}_{n,i}$, we have $Y_{n,i}(U_{n,i}) = Y_{n,i}(B_nZ_{n,i}) + X'_{n,i}\nabla Y_{n,i}(B_nZ_{n,i} + t_{n,i}X_{n,i})$. We define $\varphi_{n,i} = \nabla Y_{n,i}(v_{n,i})$, where $v_{n,i} = B_nZ_{n,i} + \bar{t}_{n,i}X_{n,i}$ and $\bar{t}_{n,i} = \sup \mathcal{T}_{n,i}$. Now, $E[X_{n,i}Y_{n,i}] = E[X_{n,i}]Y_{n,i}(B_nZ_{n,i}) + E[X'_{n,i}\varphi_{n,i}] = E[X'_{n,i}\varphi_{n,i}]$. The rest of the proof is as for Theorem 1. □

The following lemma will be useful for establishing asymptotic normality.

Lemma A.1. *Let $V_{n,i}$ is a row-wise independent triangular array and $\mu_{n,i} = E[V_{n,i}]$. Suppose that $R_{n,1}, \ldots, R_{n,n}$ are independent of $V_{n,1}, \ldots, V_{n,n}$ and that Assumption 4 holds. Moreover, assume that*

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

is bounded for some $\delta > 0$,

$$\sum_{i=1}^{n} \mu_{n,i} = 0, \tag{A.1}$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{var}(V_{n,i}) \to \sigma^2,$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \mu_{n,i}^2 \to \kappa^2,$$

where $\sigma^2 + (1-\rho)\kappa^2 > 0$. *Then*

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{n} R_{n,i}V_{n,i} \xrightarrow{d} \mathcal{N}(0, \sigma^2 + (1-\rho)\kappa^2),$$

where $N = \sum_{i=1}^{n} R_{n,i}$.

Proof: Notice that

$$E\left[\frac{N}{n\rho_n}\right] = 1$$

and

$$\mathrm{var}\left(\frac{N}{n\rho_n}\right) = \frac{n\rho_n(1-\rho_n)}{(n\rho_n)^2} \to 0.$$

Now the continuous mapping theorem implies

$$\left(\frac{n\rho_n}{N}\right)^{1/2} \xrightarrow{p} 1.$$

As a result, it is enough to prove

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{R_{n,i}}{\sqrt{\rho_n}}V_{n,i} \to \mathcal{N}(0, \sigma^2 + (1-\rho)\kappa^2).$$

Let

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n} \left(\mathrm{var}(V_{n,i}) + (1-\rho_n)\mu_{n,i}^2\right).$$

Consider n large enough so $s_n^2 > 0$. Notice that, for $i = 1, \ldots, n$,

$$E\left[\frac{R_{n,i}V_{n,i} - \rho_n\mu_{n,i}}{s_n\sqrt{n\rho_n}}\right] = 0,$$

and

$$\begin{aligned}\operatorname{var}\left(R_{n,i}V_{n,i} - \rho_n\mu_{n,i}\right) &= \rho_n E[V_{n,i}^2] - \rho_n^2\mu_{n,i}^2 \\ &= \rho_n\left(\operatorname{var}(V_{n,i}) + (1-\rho_n)\mu_{n,i}^2\right).\end{aligned}$$

Therefore,

$$\sum_{i=1}^{n}\operatorname{var}\left(\frac{R_{n,i}V_{n,i} - \rho_n\mu_{n,i}}{s_n\sqrt{n\rho_n}}\right) = 1.$$

Using $\rho_n \leq \rho_n^{1/(2+\delta)}$, $|\mu_{n,i}|^{2+\delta} \leq E[|V_{n,i}|^{2+\delta}]$, and Minkowski's inequality, we obtain:

$$\begin{aligned}\sum_{i=1}^{n}E\left[\left|\frac{R_{n,i}V_{n,i} - \rho_n\mu_{n,i}}{s_n\sqrt{n\rho_n}}\right|^{2+\delta}\right] &\leq \frac{1}{s_n^{2+\delta}(n\rho_n)^{1+\delta/2}}\sum_{i=1}^{n}\left(\rho_n^{\frac{1}{2+\delta}}\left(E\left[|V_{n,i}|^{2+\delta}\right]\right)^{\frac{1}{2+\delta}} + \rho_n|\mu_{n,i}|\right)^{2+\delta} \\ &\leq \frac{2^{2+\delta}\rho_n}{s_n^{2+\delta}(n\rho_n)^{1+\delta/2}}\sum_{i=1}^{n}E\left[|V_{n,i}|^{2+\delta}\right] \\ &= \frac{2^{2+\delta}}{s_n^{2+\delta}(n\rho_n)^{\delta/2}}\left(\frac{1}{n}\sum_{i=1}^{n}E\left[|V_{n,i}|^{2+\delta}\right]\right) \to 0.\end{aligned}$$

Applying Liapunov's theorem (see, e.g., Davidson, 1994), we obtain

$$\sum_{i=1}^{n}\frac{R_{n,i}V_{n,i} - \rho_n\mu_{n,i}}{s_n\sqrt{n\rho_n}} \xrightarrow{d} \mathcal{N}(0,1).$$

Now, the result of the lemma follows from equation (A.1) and from $s_n/\sqrt{\sigma^2 + (1-\rho)\kappa^2} \to 1$. □

Lemma A.2. *Suppose Assumptions 3-9 hold, and let* $\Delta^{\mu} = \Delta^{\text{ehw}} - \Delta^{\text{cond}}$, $\widetilde{\varepsilon}_{n,i} = Y_{n,i} - X_{n,i}'\theta_n^{\text{causal,sample}} - X_{n,i}'\gamma_n^{\text{causal,sample}}$, *and* $\nu_{n,i} = Y_{n,i} - X_{n,i}'\theta_n^{\text{descr}} - X_{n,i}'\gamma_n^{\text{descr}}$. *Then,*

(i)

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

(ii)

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{n}R_{n,i}X_{n,i}\widetilde{\varepsilon}_{n,i} \xrightarrow{d} \mathcal{N}(0, \Delta^{\text{cond}}),$$

(iii)

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{n}R_{n,i}X_{n,i}\nu_{n,i} \xrightarrow{d} \mathcal{N}(0, (1-\rho)\Delta^{\text{ehw}}).$$

Proof of Lemma A.2: To prove (i), consider $V_{n,i} = a'X_{n,i}\varepsilon_{n,i}$ for $a \in \mathbb{R}^k$. We will verify the conditions Lemma A.1. Notice that,

$$\frac{1}{n}\sum_{i=1}^{n} E\left[|V_{n,i}|^{2+\delta}\right] \leq \frac{\|a\|^{2+\delta}}{n}\sum_{i=1}^{n} E\left[\|X_{n,i}\|^{2+\delta}\left(|Y_{n,i}| + \|X_{n,i}\|\|\theta_n\| + \|Z_{n,i}\|\|\gamma_n\|\right)^{2+\delta}\right].$$

By Minkowski's inequality and Assumption 5, the right-hand side of last equation is bounded. In addition,

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

Let $a \neq 0$. Then,

$$\frac{1}{n}\sum_{i=1}^{n}\text{var}(V_{n,i}) = a'\left(\frac{1}{n}\sum_{i=1}^{n}\text{var}\left(X_{n,i}\varepsilon_{n,i}\right)\right)a \to a'\Delta^{cond}a > 0.$$

$$\frac{1}{n}\sum_{i=1}^{n}\mu_{n,i}^2 = a'\left(\frac{1}{n}\sum_{i=1}^{n} E[X_{n,i}\varepsilon_{n,i}]E[\varepsilon_{n,i}X_{n,i}']\right)a \to a'\Delta^{\mu}a.$$

This implies

$$a'\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{n} R_{n,i}X_{n,i}\varepsilon_{n,i}\right) \xrightarrow{d} \mathcal{N}(0, a'(\Delta^{cond} + (1-\rho)\Delta^{\mu})a).$$

Using the Cramer-Wold device, this implies

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

The proofs of (ii) and (iii) are similar. □

Proof of Theorem 3 : To prove (i), notice that

$$\sum_{i=1}^{n} R_{n,i}\begin{pmatrix} X_{n,i}X_{n,i}' & X_{n,i}Z_{n,i}' \\ Z_{n,i}X_{n,i}' & Z_{n,i}Z_{n,i}' \end{pmatrix}$$

is invertible with probability approaching one. Then,

$$\begin{aligned}\begin{pmatrix} \widehat{\theta}_n \\ \widehat{\gamma}_n \end{pmatrix} &= \left(\sum_{i=1}^{n} R_{n,i}\begin{pmatrix} X_{n,i}X_{n,i}' & X_{n,i}Z_{n,i}' \\ Z_{n,i}X_{n,i}' & Z_{n,i}Z_{n,i}' \end{pmatrix}\right)^{-1}\sum_{i=1}^{n} R_{n,i}\begin{pmatrix} X_{n,i}Y_{n,i} \\ Z_{n,i}Y_{n,i} \end{pmatrix} \\ &= \begin{pmatrix} \theta_n^{causal} \\ \gamma_n^{causal} \end{pmatrix} + \left(\sum_{i=1}^{n} R_{n,i}\begin{pmatrix} X_{n,i}X_{n,i}' & X_{n,i}Z_{n,i}' \\ Z_{n,i}X_{n,i}' & Z_{n,i}Z_{n,i}' \end{pmatrix}\right)^{-1}\sum_{i=1}^{n} R_{n,i}\begin{pmatrix} X_{n,i}\varepsilon_{n,i} \\ Z_{n,i}\varepsilon_{n,i} \end{pmatrix}.\end{aligned}$$

Therefore,

$$\sqrt{N}\begin{pmatrix} \widehat{\theta}_n - \theta_n^{causal} \\ \widehat{\gamma}_n - \gamma_n^{causal} \end{pmatrix} = \left(\frac{1}{N}\sum_{i=1}^{n} R_{n,i}\begin{pmatrix} X_{n,i}X_{n,i}' & X_{n,i}Z_{n,i}' \\ Z_{n,i}X_{n,i}' & Z_{n,i}Z_{n,i}' \end{pmatrix}\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{n} R_{n,i}\begin{pmatrix} X_{n,i}\varepsilon_{n,i} \\ Z_{n,i}\varepsilon_{n,i} \end{pmatrix}$$

$$= \begin{pmatrix} \Omega_n^{XX} & \Omega_n^{XZ} \\ \Omega_n^{ZX} & \Omega_n^{ZZ} \end{pmatrix}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{n} R_{n,i} \begin{pmatrix} X_{n,i}\varepsilon_{n,i} \\ Z_{n,i}\varepsilon_{n,i} \end{pmatrix} + r_n,$$

where

$$r_n = \left[\begin{pmatrix} \widetilde{W}_n^{XX} & \widetilde{W}_n^{XZ} \\ \widetilde{W}_n^{ZX} & \widetilde{W}_n^{ZZ} \end{pmatrix}^{-1} - \begin{pmatrix} \Omega_n^{XX} & \Omega_n^{XZ} \\ \Omega_n^{ZX} & \Omega_n^{ZZ} \end{pmatrix}^{-1} \right] \frac{1}{\sqrt{N}} \sum_{i=1}^{n} R_{n,i} \begin{pmatrix} X_{n,i}\varepsilon_{n,i} \\ Z_{n,i}\varepsilon_{n,i} \end{pmatrix}.$$

Because (i) $\Omega_n^{XZ} = 0$, (ii) the first term of r_n is $o_p(1)$, and (iii) $(1/\sqrt{N})\sum_{i=1}^{n} R_{n,i}X_{n,i}\varepsilon_{n,i}$ is $O_p(1)$ (under the conditions stated above), $(1/\sqrt{N})\sum_{i=1}^{n} R_{n,i}Z_{n,i}\varepsilon_{n,i} = O_p(1)$ would imply

$$\sqrt{N}(\widehat{\theta}_n - \theta_n^{causal}) = \left(\Omega_n^{XX}\right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{n} R_{n,i}X_{n,i}\varepsilon_{n,i} + o_p(1).$$

By Markov's inequality, it is enough to show that the second moment of $(1/\sqrt{N})\sum_{i=1}^{n} R_{n,i}Z_{n,i}\varepsilon_{n,i}$ is uniformly bounded. As before, we will assign an arbitrary value of zero to this quantity for the case $N = 0$. Therefore,

$$E\left[\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{n} R_{n,i}Z_{n,i}\varepsilon_{n,i} \right)^2 \right] = \sum_{i=1}^{n} E\left[\frac{R_{n,i}}{N} \,\middle|\, N > 0 \right] Z_{n,i}E[\varepsilon_{n,i}^2]Z_{n,i}'.$$

Notice that

$$E\left[\frac{R_{n,i}}{N} \,\middle|\, N > 0 \right] = \sum_{m=1}^{n} \frac{m/n}{m} \frac{\Pr(N = m)}{\Pr(N > 0)} = \frac{1}{n}.$$

As a result, it suffices that

$$\frac{1}{n} \sum_{i=1}^{n} Z_{n,i}E[\varepsilon_{n,i}^2]Z_{n,i}'$$

is uniformly bounded, which is implied by Assumption 5. The proofs of (ii) and (iii) are analogous. □

Proof of Theorem 4: The result follows directly $E[X_{n,i}\varepsilon_{n,i}] = 0$. □

Proof of Lemma 2: First, notice that (with probability approaching one) Λ_n exists and it is equal to B_n. This implies,

$$\widehat{\Lambda}_n - \Lambda_n = \left(\frac{1}{N} \sum_{i=1}^{n} R_{n,i}X_{n,i}Z_{n,i}' \right) \left(\frac{1}{N} \sum_{i=1}^{n} R_{n,i}Z_{n,i}Z_{n,i}' \right)^{-1}$$

which converges to zero in probability by Lemma 1 and Assumption 6. Direct calculations yield

$$\widehat{H}_n - \widetilde{W}_n^{XX} = (\widehat{\Lambda}_n - \Lambda_n)\widetilde{W}_n^{ZZ}(\widehat{\Lambda}_n - \Lambda_n)' - \widetilde{W}_n^{XZ}(\widehat{\Lambda}_n - \Lambda_n)' - (\widehat{\Lambda}_n - \Lambda_n)\widetilde{W}_n^{XZ} \xrightarrow{p} 0.$$

Now, Lemma 1 and Assumption 6 imply $\widehat{H}_n \xrightarrow{p} H$, where H is full rank. Theorem 3 direclty implies $\widehat{\theta}_n - \theta_n^{\text{causal}} \xrightarrow{p} 0$. $\widehat{\gamma}_n - \gamma_n^{\text{causal}} \xrightarrow{p} 0$ follows from Lemma 1. Let

$$\breve{\Delta}_n^{\text{ehw}} = \frac{1}{N} \sum_{i=1}^{n} R_{n,i}X_{n,i}\widehat{\varepsilon}_{n,i}^2 X_{n,i}', \qquad \widetilde{\Delta}_n^{\text{ehw}} = \frac{1}{N} \sum_{i=1}^{n} R_{n,i}X_{n,i}\varepsilon_{n,i}^2 X_{n,i}',$$

and

$$\Delta_n^{\text{ehw}} = \frac{1}{n}\sum_{i=1}^{n} E[X_{n,i}\varepsilon_{n,i}^2 X_{n,i}'].$$

Let α be a multi-index of dimension equal to the length of $T_{n,i} = (Y_{n,i} : X_{n,i}' : Z_{n,i}')$. In addition, let

$$\widetilde{T}_n^{\alpha} = \frac{1}{N}\sum_{i=1}^{n} \widetilde{T}_{n,i}^{\alpha} = \frac{1}{N}\sum_{i=1}^{n} R_{n,i} T_{n,i}^{\alpha},$$

and

$$\Psi_n^{\alpha} = \frac{1}{n}\sum_{i=1}^{n} E[W_{n,i}^{\alpha}].$$

Using the same argument as in the proof of Lemma 1 and given that Assumption 5 holds with $\delta = 4$, it follows that $\widetilde{T}_n^{\alpha} - \Psi_n^{\alpha} \xrightarrow{p} 0$ for $|\alpha| \leq 4$. This result directly implies $\widetilde{\Delta}_n^{\text{ehw}} - \Delta_n^{\text{ehw}} \xrightarrow{p} 0$. By the same argument plus convergence of $\widehat{\theta}_n$ and $\widehat{\gamma}_n$, it follows that $\widehat{\Delta}_n^{\text{ehw}} - \breve{\Delta}_n^{\text{ehw}} \xrightarrow{p} 0$ and $\breve{\Delta}_n^{\text{ehw}} - \widetilde{\Delta}_n^{\text{ehw}} \xrightarrow{p} 0$. Now, the result follows from $\widehat{\Delta}_n^{\text{ehw}} - \Delta^{\text{ehw}} = (\widehat{\Delta}_n^{\text{ehw}} - \breve{\Delta}_n^{\text{ehw}}) + (\breve{\Delta}_n^{\text{ehw}} - \widetilde{\Delta}_n^{\text{ehw}}) + (\widetilde{\Delta}_n^{\text{ehw}} - \Delta_n^{\text{ehw}}) + (\Delta_n^{\text{ehw}} - \Delta^{\text{ehw}}) \xrightarrow{p} 0$, where the last difference goes to zero by Assumption 9. □

Proof of Lemma 3: Notice that,

$$\widehat{\Delta}_n^{Z} = \widehat{\Delta}_n^{\text{ehw}} - \widehat{\Delta}_n^{\text{proj}}, \quad \text{where} \quad \widehat{\Delta}_n^{\text{proj}} = \frac{1}{N}\sum_{i=1}^{n} R_{n,i}\widehat{G}_n Z_{n,i} Z_{n,i}' \widehat{G}_n',$$

so that $\widehat{\Delta}_n^{Z}$ is no larger than $\widehat{\Delta}_n^{\text{ehw}}$ in a matrix sense.

Let

$$G_n = \left(\frac{1}{n}\sum_{i=1}^{n} E[X_{n,i}\varepsilon_{n,i}]Z_{n,i}'\right)\left(\frac{1}{n}\sum_{i=1}^{n} Z_{n,i}Z_{n,i}'\right)^{-1},$$

be the expected value of $\widehat{G}_n$. Under the assumptions of Lemma 2 and using the same argument as in the proof of that lemma, we obtain $\widehat{G}_n - G_n \xrightarrow{p} 0$. Therefore, $\widehat{\Delta}_n^{\text{proj}} - \Delta_n^{\text{proj}} \xrightarrow{p} 0$, where

$$\Delta_n^{\text{proj}} = \frac{1}{n}\sum_{i=1}^{n} G_n Z_{n,i} Z_{n,i}' G_n'.$$

Moreover, $\widehat{\Delta}_n^{Z} - \Delta_n^{Z} \xrightarrow{p} 0$, where $\Delta_n^{Z} = \Delta_n^{\text{ehw}} - \Delta_n^{\text{proj}}$ and

$$\Delta_n^{\text{ehw}} = \frac{1}{n}\sum_{i=1}^{n} E[X_{n,i}\varepsilon_{n,i}^2 X_{n,i}'].$$

Let

$$\Delta_n^{\mu} = \frac{1}{n}\sum_{i=1}^{n} E[X_{n,i}\varepsilon_{n,i}]E[\varepsilon_{n,i}X_{n,i}'].$$

Notice that

$$\Delta_n^{\mu} - \Delta_n^{\text{proj}} = \frac{1}{n}\sum_{i=1}^{n} E[X_{n,i}\varepsilon_{n,i}]E[\varepsilon_{n,i}X_{n,i}']$$

$$- \left(\frac{1}{n} \sum_{i=1}^{n} E[X_{n,i} \varepsilon_{n,i}] Z'_{n,i} \right) \left(\frac{1}{n} \sum_{i=1}^{n} Z_{n,i} Z'_{n,i} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} Z_{n,i} E[\varepsilon_{n,i} X'_{n,i}] \right).$$

Let $\boldsymbol{A}_n$ and $\boldsymbol{D}_n$ be the matrices with i-th rows equal to $E[\varepsilon_{n,i} X'_{n,i}]/\sqrt{n}$ and $Z'_{n,i}/\sqrt{n}$, respectively. Let $\boldsymbol{I}_n$ be the identity matrix of size n. Then,

$$\Delta_n^{\mu} - \Delta_n^{\text{proj}} = \boldsymbol{A}'_n (\boldsymbol{I}_n - \boldsymbol{D}_n (\boldsymbol{D}'_n \boldsymbol{D}_n)^{-1} \boldsymbol{D}'_n) \boldsymbol{A}_n,$$

which is positive semi-definite. Because $\Delta_n^{\text{cond}} = \Delta_n^{\text{ehw}} - \Delta_n^{\mu}$, we obtain,

$$\Delta_n^{\text{cond}} \leq \Delta_n^{Z} \leq \Delta_n^{\text{ehw}}$$

where the inequalities are to be understood in a matrix sense. Now, it follow from Assumption 11 that G_n and, therefore, Δ_n^{proj} and Δ_n^{Z} have limits. Then,

$$\Delta^{\text{cond}} \leq \Delta^{Z} \leq \Delta^{\text{ehw}}$$

where Δ^{cond}, Δ^{Z}, and Δ^{ehw} are the limits of Δ_n^{cond}, Δ_n^{Z}, and Δ_n^{ehw}, respectively. □

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2014). Finite population causal standard errors. Technical report, National Bureau of Economic Research.

Abadie, A. and G. W. Imbens (2008). Estimation of the conditional variance in paired experiments. *Annales d'Economie et de Statistique*, 175–187.

Abadie, A., G. W. Imbens, and F. Zheng (2014). Inference for misspecified models with fixed regressors. *Journal of the American Statistical Association 109*(508), 1601–1614.

Angrist, J. and S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricists' Companion*. Princeton University Press.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica 66*(2), 249–288.

Aronow, P. M. and C. Samii (2016). Does regression produce representative estimates of causal effects? *American Journal of Political Science 60*(1), 250–267.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. Advanced Texts in Econometrics. Oxford University Press.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424–455.

Efron, B. (1987). *The Jackknife, the Bootstrap, and Other Resampling Plans*, Volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 59–82.

Fogarty, C. B. (2016). Regression assisted inference for the average treatment effect in paired experiments. *arXiv preprint arXiv:1612.05179*.

Freedman, D. (2008a). On regression adjustmens to experimental data. *Advances in Applied Mathematics 40*(2), 180–193.

Freedman, D. A. (2008b). On regression adjustments in experiments with several treatments. *The annals of applied statistics 2*(1), 176–196.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–970.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 221–233.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Lin, W. (2013). Agnostic notes on regression adjustments for experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics 7*(1), 295–318.

MacKinnon, J. G. and H. White (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics 29*(3), 305–325.

Manski, C. F. (2013). *Public policy in an uncertain world: analysis and decisions*. Harvard University Press.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. Section 9. Reprinted and translated in *Statistical Science*, 5, 465–480, 1990.

Rosenbaum, P. R. (2002). Observational studies. In *Observational Studies*. Springer.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Samii, C. and P. M. Aronow (2012). On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statistics & Probability Letters 82*(2), 365–370.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Słoczyński, T. (2017). A general weighted average representation of the ordinary and two-stage least squares estimands.

White, H. (1980a). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica 48*(1), 817–838.

White, H. (1980b). Using least squares to approximate unknown regression functions. *International Economic Review 21*(1), 149–170.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.