Stanford | Institute for Economic Policy Research (SIEPR)

Finite Population Causal Standard Errors

Alberto Abadie

Susan Athey

Guido W. Imbens

Jeffrey M. Wooldridge

July, 2014

Working Paper No. 17-036

Finite Population Causal Standard Errors*

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

Current version July 2014 – First version September 2013

Abstract

When a researcher estimates the parameters of a regression function using information on all 50 states in the United States, or information on all visits to a website, what is the interpretation of the standard errors? Researchers typically report standard errors that are designed to capture sampling variation, based on viewing the data as a random sample drawn from a large population of interest, even in applications where it is difficult to articulate what that population of interest is and how it differs from the sample. In this paper we explore alternative interpretations for the uncertainty associated with regression estimates. As a leading example we focus on the case where some parameters of the regression function are intended to capture causal effects. We derive standard errors for causal effects using a generalization of randomization inference. Intuitively, these standard errors capture the fact that even if we observe outcomes for all units in the population of interest, there are for each unit missing potential outcomes for the treatment levels the unit was not exposed to. We show that our randomization-based standard errors in general are smaller than the conventional robust standard errors, and provide conditions under which they agree with them. More generally, correct statistical inference requires precise characterizations of the population of interest, the parameters that we aim to estimate within such population, and the sampling process. Estimation of causal parameters is one example where appropriate inferential methods may differ from conventional practice, but there are others.

*We are grateful for comments by Daron Acemoglu, Joshua Angrist, Matias Cattaneo, Jim Poterba, Bas Werker, and seminar participants at Microsoft Research, Michigan, MIT, Stanford, Princeton, NYU, Columbia, Tilburg University, the Tinbergen Institute, and University College London, and especially for discussions with Gary Chamberlain.

†Professor of Public Policy, John F. Kennedy School of Government, Harvard University, and NBER, alberto_abadie@harvard.edu.

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

§Professor of Economics, and Graduate School of Business Trust Faculty Fellow 2013-2014, Graduate School of Business, Stanford University, and NBER, imbens@stanford.edu.

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

1 Introduction

In many empirical studies in economics, researchers specify a parametric relation between observable variables in a population of interest. They then proceed to estimate and do inference for the parameters of this relation. Point estimates are based on matching the relation between the variables in the population to the relation observed in the sample, following what Goldberger (1968) and Manski (1988) call the "analogy principle." In the simplest setting with an observed outcome and no covariates the parameter of interest might simply be the population mean, estimated by the sample average. Given a single covariate, the parameters of interest might consist of the slope and intercept of the best linear predictor for the relationship between the outcome and the covariate. The estimated value of a slope parameter might be used to answer an economics question such as, what is the average impact of a change in the minimum wage on employment? Or, what will be the average (over markets) of the increase in demand if a firm lowers its posted price? A common hypothesis to test is that the population value of the slope parameter of the best linear predictor is equal to zero.

The textbook approach to conducting inference in such contexts relies on the assumptions that (i) the observed units are a random sample from a large population, and (ii) the parameters in this population are the objects of interest. Uncertainty regarding the parameters of interest arises from sampling variation, due to the difference between the sample and the population. A 95% confidence interval has the interpretation that if one repeatedly draws new random samples from this population and construct new confidence intervals for each sample, the estimand should be contained in the confidence interval 95% of the time. In many cases this random sampling perspective is attractive. If one analyzes individual-level data from the Current Population Survey, the Panel Study of Income Dynamics, the 1% public use sample from the Census, or other public use surveys, it is clear that the sample analyzed is only a small subset of the population of interest. However, in this paper we argue that there are other settings where there is no population such that the sample can be viewed as small relative to that population, randomly drawn from it, and when the estimand is the population value of that parameter. For example, suppose that the units are all fifty states of the United States, all the countries in the world, or all visits to a website. If we observe a cross-section of outcomes at a single point in time and ask how the average outcome varies with attributes of the units, the answer is a quantity that is known with certainty. For example, the difference in average outcome between coastal

and inland states for the observed year is known: the sample average difference is equal to the population average difference. Thus the standard error on the estimate of the difference should be zero. However, without exception researchers report positive standard errors in such settings. More precisely, researchers typically report standard errors using formulas that are formally justified by the assumption that the sample is drawn randomly from an infinite population. The theme in this paper is that this random-sampling-from-a-large-population assumption is often not the natural one for the problem at hand, and that there are other, more natural interpretations of the uncertainty in the estimates.

The general perspective we take is that statistics is fundamentally about drawing inferences with incomplete data. If the researcher sees all relevant data, there is no need for inference, since any question can be answered by simply doing calculations on the data. Outside of this polar case, it is important to be precise in what sense the data are incomplete. Often we can consider a population of units and a set of possible states of the world. There is a set of variables that takes on different values for each unit depending on the state of the world. The sampling scheme tells us how units and states of the world are chosen to form a sample, and what variables are observed, and what repeated sampling perspective may be reasonable.

Although there are many settings to consider, in the current paper we focus on the specific case where the state of the world corresponds to the level of a *causal* variable for each unit, e.g., a government regulation or a price set by a firm. The question of interest concerns the average causal effect of the variable: for example, the difference between the average outcome if (counterfactually) all units in the population are treated, and the average outcome if (counterfactually) all units in the population are not. Note that we will never observe the values for all variables of interest, because by definition we observe each physical unit at most once, either in the state where it is treated or the state where it is not, with the value of the outcome in the other state missing. Questions about causal effects can be contrasted with *descriptive* or *predictive* questions. An example of a descriptive estimand is the difference between the average outcome for countries with one set of institutions and the average outcome for countries with a different set of institutions. Although researchers often focus on causal effects in discussions about the interpretation of findings, standard practice does not distinguish between descriptive and causal estimands when conducting estimation and inference. In this paper, we show that this distinction matters. Although the distinction between descriptive estimands and causal estimands is typically not important for estimation under exogeneity assumptions, and is also

immaterial for inference if population size is large relative to the sample size, the distinction between causal and descriptive estimands matters for inference if the sample size is more than a negligible fraction of the population size. As a result the researcher should explicitly distinguish between regressors that are potential causes and those that are fixed attributes.

Although this focus on causal estimands is rarely made explicit in regression settings, it does have a long tradition in randomized experiments. In that case the natural estimator for the average causal effect is the difference in average outcomes by treatment status. In the setting where the sample and population coincide, Neyman (1923) derived the variance for this estimator and proposed a conservative estimator for this variance. The results in the current paper can be thought of as extending Neyman's analysis to general regression estimators in observational studies. Our formal analysis allows for discrete or continuous treatments and for the presence of attributes that are potentially correlated with the treatments. Thus, our analysis applies to a wide range of regression models that might be used to answer questions about the impact of government programs or about counterfactual effects of business policy changes, such as changes in prices, quality, or advertising about a product. We make four formal contributions. First, the main contribution of the study 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 (1980ab, 1982), EHW from hereon, in two directions. We allow the population to be finite, and we allow the regressors to be potential causes or attributes, or a combination of both. We take account of both the uncertainty arising from random sampling and the uncertainty arising from conditional randomization of the potential causes. This contribution can also be viewed as generalizing results from Neyman (1923) to settings with multiple linear regression estimators with both treatments and attributes that are possibly correlated. In the second contribution, we show that in general, as in the special, single-binary-covariate case that Neyman considers, the conventional EHW robust standard errors are conservative for the standard errors for the estimators for the causal parameters. Third, we show that in the case with attributes that are correlated with the treatments one can generally improve on the EHW variance estimator if the population is finite, and we propose estimators for the standard errors that are generally smaller than the EHW standard errors. Fourth, we show that in a few special cases the EHW standard errors are consistent for the true standard deviation of the least squares estimator.

By using a randomization inference approach the current paper builds on a large litera-

ture going back to Fisher (1935) and Neyman (1923). The early literature focused on settings with randomized assignment without additional covariates. See Rosenbaum (1995) and Imbens and Rubin (2014) for textbook discussions. More recent studies analyze regression methods with additional covariates under the randomization distribution in randomized experiments, e.g., Freedman (2008ab), Lin (2013), Samii and Aronow (2012), and Schochet (2010). For applications of randomization inference in observational studies see Rosenbaum (2002), Abadie, Diamond and Hainmueller (2010), Imbens and Rosenbaum (2005), Frandsen (2012), Bertrand, Duflo, and Mullainathan (2004) and Barrios, Diamond, Imbens and Kolesar (2012). In most of these studies, the assignment of the covariates is assumed to be completely random, as in a randomized experiment. Rosenbaum (2002) allows for dependence between the assignment mechanism and the attributes by assuming a logit model for the conditional probability of assignment to a binary treatment. He estimates the effects of interest by minimizing test statistics based on conditional randomization. In the current paper, we allow explicitly for general dependence of the assignment mechanism of potential causes (discrete or continuous) on the fixed attributes (discrete or continuous) of the units, thus making the methods applicable to general regression settings.

Beyond questions of causality in a given cross-section, there are other kinds of questions one could ask where the definition of the population and the sampling scheme look different; for example, we might consider the population as consisting of units in a variety of potential states of the world, where the state of the world affects outcomes through an unobservable variable. For example, we could think of a population where a member consists of a country with different realizations of weather, where weather is not in the observed data, and we wish to draw inferences about what the impact of regulation on country-level outcomes would be in a future year with different realizations of weather outcomes. We present some thoughts on this type of question in Section 6.

2 Three Examples

In this section we set the stage for the problems discussed in the current paper by introducing three simple examples for which the results are well known from either the finite population survey literature (*e.g.,* Cochran, 1977; Kish, 1995), or the causal literature (*e.g.,* Neyman, 1923; Rubin, 1974; Holland, 1986; Imbens and Wooldridge, 2008; Imbens and Rubin, 2014).

Juxtaposing these examples will provide the motivation for, and insight into, the problems we study in the current paper.

2.1 Inference for a Finite Population Mean with Random Sampling

Suppose we have a population of size M, where M may be small, large, or infinite. In the first example we focus on the simplest setting where the regression model includes only an intercept. Associated with each unit i is a non-stochastic variable Y_i, with $\boldsymbol{Y}_M$ denoting the $M-$vector with i^{th} element Y_i. The target, or estimand, is the population mean of Y_i,

$$\mu_M = \overline{Y}_M^{\text{pop}} = \frac{1}{M}\sum_{i=1}^{M} Y_i.$$

We index μ_M by the population size M because for some of the formal results we consider sequences of experiments with populations of increasing size. In that case we make assumptions that ensure that the sequence $\{\mu_M : M = 1, 2 \ldots\}$ converges to a finite constant μ, but allow for the possibility that the population mean varies over the sequence. The dual notation for the same object, μ_M and $\overline{Y}_M^{\text{pop}}$, captures the dual aspects of this quantity: on the one hand it is a population quantity, for which it is common to use Greek symbols. On the other hand, because the population is finite, it is a simple average, and the $\overline{Y}_M^{\text{pop}}$ notation shows the connection to averages. To make the example specific, one can think of the units being the 50 states ($M = 50$), and Y_i being state-level average earnings.

We do not necessarily observe all units in this population. Let W_i be a binary variable indicating whether we observe Y_i (if $W_i = 1$) or not (if $W_i = 0$), with $\boldsymbol{W}_M$ the M-vector with i^{th} element equal to W_i, and $N = \sum_{i=1}^{M} W_i$ the sample size. We let $\{\rho_M\}_{M=1,2,\ldots}$ be a sequence of sampling probabilities, one for each population size M, where $\rho_M \in (0, 1)$. If the sequence $\{\rho_M\}_{M=1,2,\ldots}$ has a limit, we denote it s limit by ρ. We make the following assumption about the sampling process.

Assumption 1. (Random Sampling without Replacement) *Given the sequence of sampling probabilities* $\{\rho_M\}_{M=1,2,\ldots}$,

$$\text{pr}\left(\boldsymbol{W}_M = \boldsymbol{w}\right) = \rho_M^{\sum_{i=1}^{M} w_i} \cdot \left(1 - \rho_M\right)^{M - \sum_{i=1}^{M} w_i},$$

for all $\boldsymbol{w}$ *with i-th element* $w_i \in \{0, 1\}$*, and all* M*.*

This sampling scheme makes the sample size N random. An alternative is to draw a random sample of fixed size. Here we focus on the case with a random sample size in order to allow for the generalizations we consider later. Often the sample is much smaller than the population but it may be that the sample coincides with the population.

The natural estimator for the population average μ_M is the sample average:

$$\hat{\mu}_M = \overline{Y}_M^{\text{sample}} = \frac{1}{N} \sum_{i=1}^{M} W_i \cdot Y_i.$$

To be formal, let us define $\hat{\mu}_M = 0$ if $N = 0$, so $\hat{\mu}_M$ is always defined. Conditional on $N > 0$ this estimator is unbiased for the population average μ_M:

$$\mathbb{E}_{\boldsymbol{W}} \left[\hat{\mu}_M \middle| N > 0 \right] = \mathbb{E}_{\boldsymbol{W}} \left[\overline{Y}_M^{\text{sample}} \middle| N > 0 \right] = \mu_M.$$

The subscript $\boldsymbol{W}$ for the expectations operator (and later for the variance operator) indicates that this expectation is over the distribution generated by the randomness in the vector of sampling indicators $\boldsymbol{W}_M$: the M-vector $\boldsymbol{Y}_M$ is fixed. We are interested in the variance of the estimator $\hat{\mu}_M$ conditional on N:

$$\mathbb{V}_{\boldsymbol{W}} \left(\hat{\mu}_M \middle| N \right) = \mathbb{E}_{\boldsymbol{W}} \left[\left(\hat{\mu}_M - \mu_M \right)^2 \middle| N \right] = \mathbb{E}_{\boldsymbol{W}} \left[\left(\overline{Y}_M^{\text{sample}} - \overline{Y}_M^{\text{pop}} \right)^2 \middle| N \right].$$

Because we condition on N this variance is itself a random variable. It is also useful to define the normalized variance, that is, the variance normalized by the sample size N:

$$\mathbb{V}^{\text{norm}} \left(\hat{\mu}_M \right) = N \cdot \mathbb{V}_{\boldsymbol{W}} \left(\hat{\mu}_M \middle| N \right),$$

which again is a random variable. Also define

$$\sigma_M^2 = \frac{1}{M-1} \sum_{i=1}^{M} \left(Y_i - \overline{Y}^{\text{pop}} \right)^2,$$

which we refer to as the population variance (note that, in contrast to some definitions, we divide by $M - 1$ rather than M).

Here we state a slight modification of a well-known result from the survey sampling literature. The case with a fixed sample size can be found in various places in the survey sampling literature, such as Cochran (1977) and Kish (1995). Deaton (1997) also covers the result. We provide a proof because of the slight modification and because the basic argument is used in subsequent results.

Lemma 1. (Exact Variance under Random Sampling) *Suppose Assumption 1 holds. Then*

$$\mathbb{V}_{\boldsymbol{W}}\left(\hat{\mu}_M \middle| N, N>0\right) = \frac{\sigma_M^2}{N} \cdot \left(1 - \frac{N}{M}\right).$$

All proofs are in the appendix.

If the sample is close in size to the population, then the variance of the sample average as an estimator of the population average will be close to zero. The adjustment factor for the finite population, $1 - N/M$, is proportional to one minus the ratio of the sample and population size. It is rare to see this adjustment factor used in empirical studies in economics.

For the next result we rely on assumptions about sequences of populations with increasing size, indexed by the population size M. These sequences are not stochastic. We assume that the first and second moments of the population outcomes converge as the population size grows. Let $\mu_{k,M}$ be the k^{th} population moment of Y_i, $\mu_{k,M} = \sum_{i=1}^{M} Y_i^k / M$.

Assumption 2. (Sequence of Populations) *For $k = 1, 2$, and some constants μ_1, μ_2,*

$$\lim_{M\to\infty} \mu_{k,M} = \mu_k.$$

Define $\sigma^2 = \mu_2 - \mu_1^2$. We will also rely on the following assumptions on the sampling rate.

Assumption 3. (Sampling Rate) *The sequence of sampling rates ρ_M satisfies*

$$M \cdot \rho_M \to \infty, \qquad \text{and} \quad \rho_M \to \rho \in [0, 1].$$

The first part of the assumption guarantees that as the population size diverges, the (random) sample size also diverges. The second part of the assumption allows for the possibility that asymptotically the sample size is a neglible fraction of the population size.

Lemma 2. (Variance in Large Populations) *Suppose Assumptions 1-3 hold. Then: (i)*

$$\mathbb{V}_{\boldsymbol{W}}(\hat{\mu}_M|N) - \frac{\sigma^2}{N} = O_p((\rho_M \cdot M)^{-1}),$$

(where σ^2/N is ∞ if $N = 0$), and (ii), as $M \to \infty$,

$$\mathbb{V}^{\text{norm}}(\hat{\mu}_M) \xrightarrow{p} \sigma^2 \cdot (1 - \rho).$$

In particular, if $\rho = 0$, the normalized variance converges to σ^2, corresponding to the conventional result for the normalized variance.

2.2 Inference for the Difference of Two Means with Random Sampling from a Finite Population

Now suppose we are interested in the difference between two population means, say the difference in state-level average earnings for coastal and landlocked states for the 50 states in the United States. We have to be careful, because if we draw a relatively small, completely random, sample there may be no coastal or landlocked states in the sample, but the result is essentially still the same: as N approaches M, the variance of the standard estimator for the difference in average earnings goes to zero, even after normalizing by the sample size.

Let $X_i \in \{\text{coast}, \text{land}\}$ denote the geographical status of state i. Define, for $x = \text{coast}, \text{land}$, the population size $M_x = \sum_{i=1}^{M} \mathbf{1}_{X_i=x}$, and the population averages and variances

$$\mu_{x,M} = \overline{Y}^{\text{pop}}_{x,M} = \frac{1}{M_x} \sum_{i:X_i=x} Y_i, \qquad \text{and} \quad \sigma^2_{x,M} = \frac{1}{M_x - 1} \sum_{i:X_i=x} \left(Y_i - \overline{Y}^{\text{pop}}_{x,M}\right)^2.$$

The estimand is the difference in the two population means,

$$\theta_M = \overline{Y}^{\text{pop}}_{\text{coast},M} - \overline{Y}^{\text{pop}}_{\text{land},M},$$

and the natural estimator for θ_M is the difference in sample averages by state type,

$$\hat{\theta}_M = \overline{Y}^{\text{sample}}_{\text{coast}} - \overline{Y}^{\text{sample}}_{\text{land}},$$

where the averages of observed outcomes and sample sizes by type are

$$\overline{Y}^{\text{sample}}_{x} = \frac{1}{N_x} \sum_{i:X_i=x} W_i \cdot Y_i, \qquad \text{and} \quad N_x = \sum_{i:X_i=x} W_i,$$

for $x = \text{coast}, \text{land}$. The estimator $\hat{\theta}_M$ can also be thought of as the least squares estimator for θ based on minimizing

$$\arg\min_{\gamma,\theta} \sum_{i=1}^{M} W_i \cdot \left(Y_i - \gamma - \theta \cdot \mathbf{1}_{X_i=\text{coast}}\right)^2.$$

The extension of part (i) of Lemma 1 to this case is fairly immediate. Again the outcomes Y_i are viewed as fixed quantities. So are the attributes X_i, with the only stochastic component the vector $\boldsymbol{W}_M$. We condition on N_{coast} and N_{land} being positive.

Lemma 3. (Random Sampling and Regression) *Suppose Assumption 1 holds. Then*

$$\mathbb{V}_{\boldsymbol{W}}\left(\hat{\theta}\Big|\, N_{\text{land}}, N_{\text{coast}}, N_{\text{land}} > 0, N_{\text{coast}} > 0\right) = \frac{\sigma^2_{\text{coast},M}}{N_{\text{coast}}} \cdot \left(1 - \frac{N_{\text{coast}}}{M_{\text{coast}}}\right) + \frac{\sigma^2_{\text{land},M}}{N_{\text{land}}} \cdot \left(1 - \frac{N_{\text{land}}}{M_{\text{land}}}\right).$$

Again, as in Lemma 1, as the sample size approaches the population size, for a fixed population, the variance converges to zero. In the special case where the two sampled fractions are the same, $N_{\text{coastal}}/M_{\text{coastal}} = N_{\text{land}}/M_{\text{land}} = \rho$, the adjustment relative to the conventional variance is again simply the factor $1 - \rho$, one minus the sample size over the population size.

2.3 Inference for the Difference in Means given Random Assignment

This is the most of important of the three examples, and the one where many (but not all) of the issues that are central in the paper are present. Again it is a case with a single binary regressor. However, the nature of the regressor is conceptually different. To make the discussion specific, suppose the binary indicator or regressor is an indicator for the state having a minimum wage higher than the federal minimum wage, so $X_i \in \{\text{low}, \text{high}\}$. One possibility is to view this example as isomorphic to the previous example. This would imply that for a fixed population size the variance would go to zero as the sample size approaches the population size. However, we take a different approach to this problem that leads to a variance that remains positive even if the sample is identical to the population. The key to this approach is the view that this regressor is *not* a fixed attribute or characteristic of each state, but instead is a *potential cause*. The regressor takes on a particular value for each state in our sample, but its value could have been different. For example, in the real world Massachusetts has a state minimum wage that exceeds the federal one. We are interested in the comparison of the outcome, say state-level earnings, that was observed, and the counterfactual outcome that would have been observed had Massachusetts not had a state minimum wage that exceeded the federal one. Formally, using the Rubin causal model or potential outcome framework (Neyman, 1923; Rubin, 1974; Holland, 1986; Imbens and Rubin, 2014), we postulate the existence of two potential outcomes for each state, denoted by $Y_i(\text{low})$ and $Y_i(\text{high})$, for earnings without and with a state minimum wage, with Y_i the outcome corresponding to the actual or prevailing minimum wage:

$$Y_i = Y_i(X_i) = \begin{cases} Y_i(\text{high}) & \text{if } X_i = \text{high}, \\ Y_i(\text{low}) & \text{otherwise.} \end{cases}$$

It is important that these potential outcomes $(Y_i(\text{low}), Y_i(\text{high}))$ are well defined for each unit (the 50 states in our example), irrespective of whether that state has a minimum wage higher than the federal one or not. Let $\boldsymbol{Y}_M$, and $\boldsymbol{X}_M$ be the M-vectors with ith element equal to Y_i, and X_i respectively.

We now define two distinct estimands. The first is the population average causal effect of the state minimum wage, defined as

$$\theta_M^{\text{causal}} = \frac{1}{M}\sum_{i=1}^{M}\Big(Y_i(\text{high}) - Y_i(\text{low})\Big). \tag{2.1}$$

We disinguish this causal estimand from the descriptive or predictive difference in population averages by minimum wage,

$$\theta_M^{\text{descr}} = \frac{1}{M_{\text{high}}}\sum_{i:X_i=\text{high}} Y_i - \frac{1}{M_{\text{low}}}\sum_{i:X_i=\text{low}} Y_i. \tag{2.2}$$

It is the difference between the two estimands, θ^{causal} and θ^{descr}, that is at the core of our paper. First, we argue that although researchers are often interested in causal rather than descriptive estimands, this distinction is not often made explicit. However, many textbook discussions formally define estimands in a way that corresponds to descriptive estimands.[1] Second, we show that in settings where the sample size is of the same order of magnitude as the population size, the distinction between the causal and descriptive estimands matters. In such settings the researcher therefore needs to be explicit about the causal or descriptive nature of the estimand.

Let us start with the first point, the relative interest in the two estimands, θ_M^{causal} and θ_M^{descr}. Consider a setting where a key regressor is a state regulation. The descriptive estimand is the average difference in outcomes between states with and states without the regulation. The causal estimand is the average difference, over all states, of the outcome with and without that regulation for that state. We would argue that in such settings the causal estimand is of more interest than the descriptive estimand.

[1]For example, Goldberger (1968) writes "Regression analysis is essentially concerned with estimation of such a population regression function on the basis of a sample of observations drawn from the joint probability distribution of Y_i, X_i." (Goldberger, 1968, p. 3). Wooldridge (2002) writes: "More precisely, we assume that (1) a population model has been specified and (2) an independent identically distributed (i.i.d.) sample can be drawn from the population." (Wooldridge, 2002 p. 5). Angrist and Pischke (2008) write: "We therefore use samples to make inferences about populations" (Angrist and Pishke, 2008, p. 30). Gelman and Hill (2007) write: "Statistical inference is used to learn from incomplete or imperfect data. ... In the *sampling model* we are interested in learning some characteristic of a population ... which we must estimate from a sample, or subset, of the population. (Gelman and Hill, 2007).

Now let us study the statistical properties of the difference between the two estimands. We assume random assignment of the binary covariate X_i:

Assumption 4. (RANDOM ASSIGNMENT) *For some sequence* $\{q_M : M = 1, 2, \ldots\}$, *with* $q_M \in (0, 1)$,

$$\text{pr}\left(\boldsymbol{X} = \boldsymbol{x}\right) = q_M^{\sum_{i=1}^M 1_{x_i=\text{high}}} \cdot (1 - q_M)^{M - \sum_{i=1}^M 1_{x_i=\text{low}}},$$

for all M*-vectors* $\boldsymbol{x}$ *with* $x_i \in \{\text{low}, \text{high}\}$, *and all* M.

In the context of the example with the state minimum wage, the assumption requires that whether a state has a state minimum wage exceeding the federal wage is unrelated to the potential outcomes. This assumption, and similar ones in other cases, is arguably unrealistic, outside of randomized experiments. Often such an assumption is more plausible within homogenous subpopulations defined by observable attributes of the units. This is the motivation for including additional covariates in the specification of the regression function, and we consider such settings in the next section. For expositional purposes we proceed in this section with the simpler setting.

To formalize the relation between θ_M^{descr} and θ_M^{causal} we introduce notation for the means of the two potential outcomes, for $x = \text{low}, \text{high}$, over the entire population and by treatment status:

$$\overline{Y}_M^{\text{pop}}(x) = \frac{1}{M}\sum_{i=1}^M Y_i(x), \qquad \text{and} \quad \overline{Y}_{x,M}^{\text{pop}} = \frac{1}{M_x}\sum_{i:X_i=x} Y_i(x),$$

where, as before, $M_x = \sum_{i=1}^M \mathbf{1}_{X_i=x}$ is the population size by treatment group. Note that because X_i is a random variable, M_{high} and M_{low} are random variables too. Now we can write the two estimands as

$$\theta_M^{\text{causal}} = \overline{Y}_M^{\text{pop}}(\text{high}) - \overline{Y}_M^{\text{pop}}(\text{low}), \qquad \text{and} \quad \theta^{\text{descr}} = \overline{Y}_{\text{high},M}^{\text{pop}} - \overline{Y}_{\text{low},M}^{\text{pop}}.$$

Define the population variances of the two potential outcomes $Y_i(\text{low})$ and $Y_i(\text{high})$,

$$\sigma_M^2(x) = \frac{1}{M-1}\sum_{i=1}^M \left(Y_i(x) - \overline{Y}_M^{\text{pop}}(x)\right)^2, \qquad \text{for } x = \text{low}, \text{high},$$

and the population variance of the unit-level causal effect $Y_i(\text{high}) - Y_i(\text{low})$:

$$\sigma_M^2(\text{low}, \text{high}) = \frac{1}{M-1}\sum_{i=1}^M \left(Y_i(\text{high}) - Y_i(\text{low}) - \left(\overline{Y}^{\text{pop}}(\text{high}) - \overline{Y}^{\text{pop}}(\text{low})\right)\right)^2.$$

The following lemma describes the relation between the two population quantities. Note that θ_M^{causal} is a fixed quantity given the population, whereas θ_M^{descr} is a random variable because it depends on $\boldsymbol{X}_M$, which is random by Assumption 4. To stress where the randomness in θ_M^{descr} stems from, and in particular to distinguish this from the sampling variation, we use the subscript $\boldsymbol{X}$ on the expectations and variance operators here. Note that at this stage there is no sampling yet: the statements are about quantities in the population.

Lemma 4. (Causal versus Descriptive Estimands) *Suppose Assumption 4 holds. Then* (i) *the descriptive estimand is unbiased for the causal estimand,*

$$\mathbb{E}_{\boldsymbol{X}}[\theta_M^{\text{descr}}|M_{\text{high}}, M_{\text{low}} > 0, M_{\text{high}} > 0] = \theta_M^{\text{causal}},$$

and (ii)*,*

$$\begin{aligned}
&\mathbb{V}_{\boldsymbol{X}}\left(\theta_M^{\text{descr}}\middle|\, M_{\text{high}}, M_{\text{low}} > 0, M_{\text{high}} > 0\right) \\
&\qquad = \mathbb{E}_{\boldsymbol{X}}\left[\left(\theta_M^{\text{descr}} - \theta_M^{\text{causal}}\right)^2\middle|\, M_{\text{high}}, M_{\text{low}} > 0, M_{\text{high}} > 0\right] \\
&\qquad = \frac{\sigma_M^2(\text{low})}{M_{\text{low}}} + \frac{\sigma_M^2(\text{high})}{M_{\text{high}}} - \frac{\sigma_M^2(\text{low},\text{high})}{M} \geq 0.
\end{aligned}$$

These results are well-known from the causality literature, starting with Neyman (1923). See for a recent discussion and details Imbens and Rubin (2014).

Now let us generalize these results to the case where we only observe values for X_i and Y_i for a subset of the units in the population. As before in Assumption 1, we assume this is a random subset, but we strengthen Assumption 1 by assuming the sampling is random, conditional on $\boldsymbol{X}$.

Assumption 5. (Random Sampling Without Replacement) *Given the sequence of sampling probabilities* $\{\rho_M : M = 1, 2, \ldots\}$*, and conditional on* $\boldsymbol{X}_M$*,*

$$\text{pr}\left(\boldsymbol{W}_M = \boldsymbol{w}|\boldsymbol{X}_M\right) = \rho_M^{\sum_{i=1}^M w_i} \cdot (1-\rho_M)^{M-\sum_{i=1}^M w_i},$$

for all M*-vectors* $\boldsymbol{w}$ *with* i*-th element* $w_i \in \{0, 1\}$*, and all* M*.*

We focus on the properties of the same estimator as in the second example in Section 2.2,

$$\hat{\theta} = \overline{Y}_{\text{high}}^{\text{obs}} - \overline{Y}_{\text{low}}^{\text{obs}},$$

where, for $x \in \{\text{low}, \text{high}\}$,

$$\overline{Y}_x^{\text{obs}} = \frac{1}{N_x} \sum_{i:X_i=x} W_i \cdot Y_i, \qquad \text{and} \quad N_x = \sum_{i=1}^{M} W_i \cdot \mathbf{1}_{X_i=x}.$$

The following results are closely related to results in the causal literature. Some of the results rely on uncertainty from random sampling, some on uncertainty from random assignment, and some rely on both sources of uncertainty: the superscripts $\boldsymbol{W}$ and $\boldsymbol{X}$ clarify these distinctions.

Lemma 5. (Expectations and Variances for Causal and Descriptive Estimands) *Suppose that Assumptions 4 and 5 hold. Then:*

(i)

$$\mathbb{E}_{\boldsymbol{W},\boldsymbol{X}}\left[\hat{\theta}\middle| N_{\text{low}}, N_{\text{high}}, N_{\text{low}} > 0, N_{\text{high}} > 0\right] = \theta_M^{\text{causal}},$$

(ii)

$$\begin{aligned}&\mathbb{V}_{\boldsymbol{W},\boldsymbol{X}}\left(\hat{\theta} - \theta_M^{\text{causal}}\middle| N_{\text{low}}, N_{\text{high}}, N_{\text{low}} > 0, N_{\text{high}} > 0\right)\\ &\quad = \frac{\sigma_M^2(\text{low})}{N_{\text{low}}} + \frac{\sigma_M^2(\text{high})}{N_{\text{high}}} - \frac{\sigma_M^2(\text{low},\text{high})}{M},\end{aligned}$$

(iii)

$$\mathbb{E}_{\boldsymbol{W}}\left[\hat{\theta}\middle| \boldsymbol{X}, M_{\text{low}}, N_{\text{low}} > 0, N_{\text{high}} > 0\right] = \theta_M^{\text{descr}},$$

(iv)

$$\begin{aligned}&\mathbb{V}_{\boldsymbol{W},\boldsymbol{X}}\left(\hat{\theta} - \theta_M^{\text{descr}}\middle| M_{\text{low}}, N_{\text{low}}, N_{\text{high}}, N_{\text{low}} > 0, N_{\text{high}} > 0\right)\\ &\quad = \frac{\sigma_M^2(\text{low})}{N_{\text{low}}} \cdot \left(1 - \frac{N_{\text{low}}}{M_{\text{low}}}\right) + \frac{\sigma_M^2(\text{high})}{N_{\text{high}}} \cdot \left(1 - \frac{N_{\text{high}}}{M_{\text{high}}}\right),\end{aligned}$$

(v)

$$\begin{aligned}&\mathbb{V}_{\boldsymbol{W},\boldsymbol{X}}\left(\hat{\theta} - \theta_M^{\text{causal}}\middle| N_{\text{low}}, N_{\text{high}}, N_{\text{low}} > 0, N_{\text{high}} > 0\right)\\ &\qquad -\mathbb{V}_{\boldsymbol{W},\boldsymbol{X}}\left(\hat{\theta} - \theta_M^{\text{descr}}\middle| N_{\text{low}}, N_{\text{high}}, N_{\text{low}} > 0, N_{\text{high}} > 0\right)\\ &= \mathbb{V}_{\boldsymbol{W},\boldsymbol{X}}\left(\theta_M^{\text{desc}} - \theta_M^{\text{causal}}\middle| N_{\text{low}}, N_{\text{high}}, N_{\text{low}} > 0, N_{\text{high}} > 0\right)\\ &= \frac{\sigma_M^2(\text{low})}{M_{\text{low}}} + \frac{\sigma_M^2(\text{high})}{M_{\text{high}}} - \frac{\sigma_M^2(\text{low},\text{high})}{M} \geq 0.\end{aligned}$$

Part (ii) of Lemma 5 is a re-statement of results in Neyman (1923). Part (iv) is essentially the same result as in Lemma 2. Parts (ii) and (iv) of the lemma, in combination with Lemma 4, imply part (v). Although part (ii) and (iv) of Lemma 5 are both known in their respective literatures, the juxtaposition of the two variances has not received much attention.

Next, we study what happens in large populations. In order to do so we need to modify Assumption 2 for the current context. First, define

$$\mu_{k,m,M} = \frac{1}{M}\sum_{i=1}^{M} Y_i^k(\text{low}) \cdot Y_i^m(\text{high}).$$

We assume that all (cross-)moments up to second order converge to finite limits.

Assumption 6. (Sequence of Populations) *For nonnegative integers k, m such that $k + m \leq 2$, and some constants $\mu_{k,m}$,*

$$\lim_{M\to\infty} \mu_{k,m,M} = \mu_{k,m}.$$

Then define $\sigma^2(\text{low}) = \mu_{2,0} - \mu_{1,0}^2$ and $\sigma^2(\text{high}) = \mu_{0,2} - \mu_{0,1}^2$, so that under Assumption 6 $\lim_{M\to\infty} \sigma_M^2(\text{low}) = \sigma^2(\text{low})$ and $\lim_{M\to\infty} \sigma_M^2(\text{high}) = \sigma^2(\text{high})$. Also define, again under Assumption 6, $\lim_{M\to\infty} \sigma_M^2(\text{low}, \text{high}) = \sigma^2(\text{low}, \text{high})$.

Define the normalized variances for the causal and descriptive estimands,

$$\mathbb{V}_{\text{causal}}^{\text{norm}} = N \cdot \mathbb{V}_{\boldsymbol{W},\boldsymbol{X}}\left(\hat{\theta} - \theta_M^{\text{causal}} \middle| N_{\text{high}}, N_{\text{low}}\right),$$

and

$$\mathbb{V}_{\text{descr}}^{\text{norm}} = N \cdot \mathbb{V}_{\boldsymbol{W},\boldsymbol{X}}\left(\hat{\theta} - \theta_M^{\text{descr}} \middle| N_{\text{high}}, N_{\text{low}}\right),$$

where the variances are zero if N_{high} or N_{low} are zero.

Assumption 7. (Limiting Assignment Rate) *The sequence of assignment rates q_M satisfies*

$$\lim_{M\to\infty} q_M = q \in (0, 1).$$

Lemma 6. (Variances for Causal and Descriptive Estimands in Large Populations) *Suppose that Assumptions 3–7 hold. Then, as $M \to \infty$, (i),*

$$\mathbb{V}_{\text{causal}}^{\text{norm}} \xrightarrow{p} \frac{\sigma^2(\text{low})}{1-q} + \frac{\sigma^2(\text{high})}{q} - \rho \cdot \sigma^2(\text{low}, \text{high}) \tag{2.3}$$

and (ii),

$$\mathbb{V}_{\text{descr}}^{\text{norm}} \xrightarrow{p} \left(\frac{\sigma^2(\text{low})}{1-q} + \frac{\sigma^2(\text{high})}{q}\right) \cdot (1-\rho). \tag{2.4}$$

Lemma 6 contains some key insights. It shows that we do not need to be concerned about the difference between the two estimands θ_M^{causal} and θ_M^{descr} in settings where the population is large relative to the sample (ρ close to zero). In that case both normalized variances are equal to

$$\lim_{M\to\infty} \mathbb{V}_{\text{causal}}^{\text{norm}}\Big|_{\rho=0} = \lim_{M\to\infty} \mathbb{V}_{\text{desc}}^{\text{norm}}\Big|_{\rho=0} = \frac{\sigma^2(\text{low})}{1-q} + \frac{\sigma^2(\text{high})}{q}.$$

It is only in settings with the ratio of the sample size to the population size non-neglible, and in particular if the sample size is equal to the population size, that there are substantial differences between the two variances.

3 The Variance of Regression Estimators when the Regression includes Attributes and Causes

In this section and the next we turn to the setting that is the main focus of the current paper. We allow for the presence of covariates of the potential cause type (say a state institution or a regulation such as the state minimum wage), which can be discrete or continuous, and vector-valued. We also allow for the presence of covariates of the attribute or characteristic type, say an indicator whether a state is landlocked or coastal, which again can be discrete or continuous or discrete, and vector-valued. We allow the potential causes and attributes to be systematically correlated in the sample, because the distribution of the potential causes differs between units.

3.1 Set Up

We denote the potential causes for unit i in population M by X_{iM}, and the attributes for unit i in population M by Z_{iM}. The vector of attributes Z_{iM} typically includes an intercept. We assume there exists a set of potential outcomes $Y_{iM}(x)$, with the realized outcome for unit i in population M equal to $Y_{iM} = Y_{iM}(X_{iM})$. We sample units from this population, with W_{iM} the binary indicator for the event that unit i in population M is sampled. We view the potential outcome functions $Y_{iM}(x)$ and the attributes Z_{iM} as deterministic, and the potential causes X_{iM} and the sampling indicator W_{iM} as stochastic. However, unlike in a randomized experiment, the potential cause X_{iM} is in general not identically distributed.

For general regression we have no finite sample results for the properties of the least squares estimator. Instead we rely on large sample arguments. We formulate assumptions on the

sequence of populations, characterized by sets of covariates or attributes $\boldsymbol{Z}_M$ and potential outcomes $\boldsymbol{Y}_M(x)$, as well as on the sequence of assignment mechanisms. To be technically precise we use a double index on all variables, whether deterministic or stochastic, to reflect the fact that the distributions general depend on the population. For the asymptotics we let the size of the population M go to infinity. We allow the sampling rate, ρ_M, to be a function of the population, allowing for $\rho_M = 1$ (the sample is the population) as well as $\rho_M \to 0$ (random sampling from a large population). In the latter case case our results agree with the standard robust Eicker-Huber-White variance results based on random sampling from an infinite population. The only stochastic component is the matrix $(\boldsymbol{X}_M, \boldsymbol{W}_M)$. The randomness in $\boldsymbol{X}_M$ generates randomness in the realized outcomes $Y_{iM} = Y_{iM}(X_{iM})$ even though the potential outcome functions $Y_{iM}(x)$ are non-stochastic.

For a given population, indexed by M, define the population moments

$$\Omega_M^{\text{pop}} = \frac{1}{M}\sum_{i=1}^{M} \begin{pmatrix} Y_{iM} \\ Z_{iM} \\ X_{iM} \end{pmatrix} \begin{pmatrix} Y_{iM} \\ Z_{iM} \\ X_{iM} \end{pmatrix}',$$

and the expected population moments, where the expectation is taken over the assignment $\boldsymbol{X}$,

$$\Omega_M^{*,\text{pop}} = \mathbb{E}_{\boldsymbol{X}}\left[\Omega_M^{\text{pop}}\right] = \mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{M}\sum_{i=1}^{M} \begin{pmatrix} Y_{iM} \\ Z_{iM} \\ X_{iM} \end{pmatrix} \begin{pmatrix} Y_{iM} \\ Z_{iM} \\ X_{iM} \end{pmatrix}'\right].$$

Also define the sample moments,

$$\Omega_M^{\text{sample}} = \frac{1}{N}\sum_{i=1}^{M} W_{iM} \cdot \begin{pmatrix} Y_{iM} \\ Z_{iM} \\ X_{iM} \end{pmatrix} \begin{pmatrix} Y_{iM} \\ Z_{iM} \\ X_{iM} \end{pmatrix}'$$

where N is the random sample size.

The partitioned versions of these three matrices will be written as

$$\Omega = \begin{pmatrix} \Omega_{YY} & \Omega_{YZ'} & \Omega_{YX'} \\ \Omega_{ZY} & \Omega_{ZZ'} & \Omega_{ZX'} \\ \Omega_{XY} & \Omega_{XZ'} & \Omega_{XX'} \end{pmatrix},$$

for $\Omega = \Omega_M^{\text{pop}}$, $\Omega_M^{*,\text{pop}}$ and Ω_M^{sample}. Below we state formal assumptions that ensure that these quantities are well-defined and finite, at least in large populations.

For a population of size M, we estimate a linear regression model

$$Y_{iM} = X_{iM}'\theta + Z_{iM}'\gamma + \varepsilon_{iM},$$

by ordinary least squares, with the estimated least squares coefficients equal to

$$(\hat{\theta}_{\text{ols}}, \hat{\gamma}_{\text{ols}}) = \arg\min_{\theta,\gamma} \sum_{i=1}^{M} W_{iM} \cdot (Y_{iM} - X'_{iM}\theta - Z'_{iM}\gamma)^2 ,$$

where W_{iM} simply selects the sample that we use in estimation. The unique solution, assuming no perfect collinearity in the sample, is

$$\begin{pmatrix} \hat{\theta}_{\text{ols}} \\ \hat{\gamma}_{\text{ols}} \end{pmatrix} = \begin{pmatrix} \Omega^{\text{sample}}_{XX,M} & \Omega^{\text{sample}}_{XZ',M} \\ \Omega^{\text{sample}}_{ZX',M} & \Omega^{\text{sample}}_{ZZ',M} \end{pmatrix}^{-1} \begin{pmatrix} \Omega^{\text{sample}}_{XY,M} \\ \Omega^{\text{sample}}_{ZY,M} \end{pmatrix}.$$

We are interested in the properties of the least squares estimator for descriptive and causal estimands.

3.2 Descriptive and Causal Estimands

We now define the descriptive and causal estimands that generalize θ^{causal}_M and θ^{descr}_M from Section 2.3. For the descriptive estimand the generalization is obvious: we are interested in the value of the least squares estimator if all units in the population are observed:

$$\begin{pmatrix} \theta^{\text{descr}}_M \\ \gamma^{\text{descr}}_M \end{pmatrix} = \begin{pmatrix} \Omega^{\text{pop}}_{XX,M} & \Omega^{\text{pop}}_{XZ',M} \\ \Omega^{\text{pop}}_{ZX',M} & \Omega^{\text{pop}}_{ZZ',M} \end{pmatrix}^{-1} \begin{pmatrix} \Omega^{\text{pop}}_{XY,M} \\ \Omega^{\text{pop}}_{ZY,M} \end{pmatrix}.$$

This estimand, even though a population quantity, is stochastic because it is a function of $\boldsymbol{X}_M = (X_{1M}, X_{2M}, ..., X_{MM})'$. For the causal estimand we look at the same expression, with expectations taken over $\boldsymbol{X}$ in both components:

$$\begin{pmatrix} \theta^{\text{causal}}_M \\ \gamma^{\text{causal}}_M \end{pmatrix} = \begin{pmatrix} \Omega^{*,\text{pop}}_{XX,M} & \Omega^{*,\text{pop}}_{XZ',M} \\ \Omega^{*,\text{pop}}_{ZX',M} & \Omega^{*,\text{pop}}_{ZZ',M} \end{pmatrix}^{-1} \begin{pmatrix} \Omega^{*,\text{pop}}_{XY,M} \\ \Omega^{*,\text{pop}}_{ZY,M} \end{pmatrix}.$$

These causal parameters are non-stochastic.

To build some insight for the definition of the causal parameters, consider the special case with the attributes consisting of an intercept only, $Z_{iM} = 1$, and a single randomly assigned binary cause, $X_{iM} \in \{0, 1\}$, the case considered in Section 2.3. In that case, let, as before, $q_M = \mathbb{E}[\sum_{i=1}^{M} X_i/M]$. Then:

$$\begin{aligned} \begin{pmatrix} \theta^{\text{causal}}_M \\ \gamma^{\text{causal}}_M \end{pmatrix} &= \begin{pmatrix} \Omega^{*,\text{pop}}_{XX,M} & \Omega^{*,\text{pop}}_{XZ',M} \\ \Omega^{*,\text{pop}}_{ZX',M} & \Omega^{*,\text{pop}}_{ZZ',M} \end{pmatrix}^{-1} \begin{pmatrix} \Omega^{*,\text{pop}}_{XY,M} \\ \Omega^{*,\text{pop}}_{ZY,M} \end{pmatrix} \\ &= \begin{pmatrix} q_M & q_M \\ q_M & 1 \end{pmatrix}^{-1} \begin{pmatrix} q_M \cdot \overline{Y}^{\text{pop}}_M(1) \\ q_M \cdot \overline{Y}^{\text{pop}}_M(1) + (1-q_M) \cdot \overline{Y}^{\text{pop}}_M(0) \end{pmatrix} \\ &= \begin{pmatrix} \overline{Y}^{\text{pop}}_M(1) - \overline{Y}^{\text{pop}}_M(0) \\ \overline{Y}^{\text{pop}}_M(0) \end{pmatrix}. \end{aligned}$$

Thus $\theta^{\text{causal}}_M = \overline{Y}^{\text{pop}}_M(1) - \overline{Y}^{\text{pop}}_M(0)$, identical to the causal estimand considered in Section 2.3.

3.3 Population Residuals

We define the population residuals, denoted by ε_{iM}, to be the residual relative to the population causal estimands,

$$\varepsilon_{iM} = Y_{iM} - X'_{iM}\theta_M^{\text{causal}} - Z'_{iM}\gamma_M^{\text{causal}}.$$

The definitions of these residuals mirrors that in conventional regression analyses, but their properties are conceptually different. For example, the residuals need not be stochastic. Consider the special case where $Y_{iM}(x) = Y_{iM}(0) + x'\theta$, the potential causes X_{iM} are randomly assigned, and there are no attributes beyond the intercept. Then $\varepsilon_{iM} = Y_{iM}(0) - \sum_{j=1}^{M} Y_{jM}(0)/M$, which is non-stochastic. In other cases the residuals may be stochastic. (To be clear, the residuals are generally non-zero even though they are non-stochastic.)

Under the assumptions we make, in particular the assumption that the X_{iM} are jointly independent (but not necessarily identically distributed), the products $X_{iM} \cdot \varepsilon_{iM}$ and $Z_{iM} \cdot \varepsilon_{iM}$, are jointly independent but not identically distributed. Most importantly, in general the expectations $\mathbb{E}_{\boldsymbol{X}}[X_{iM} \cdot \varepsilon_{iM}]$ and $\mathbb{E}_{\boldsymbol{X}}[Z_{iM} \cdot \varepsilon_{iM}]$ may vary across i, although under the assumptions stated below and the definition of the residuals, the averages of these expectations over the population are guaranteed to be zero.

3.4 Assumptions

A key feature is that we now allow for more complicated assignment mechanisms. In particular, we maintain the assumption that the X_{iM}, for $i = 1, \ldots, M$, are independent but we relax the assumption that the distributions of the X_{iM} are identical. For stating general results, where the parameters are simply defined by the limits of the expected moment matrices, we do not need to restrict the distributions of the X_{iM}. However, in the case where the regression function is correctly specified, for some purposes we restrict the distribution of X_{iM} so that it depends on the Z_{iM} and not generally on the potential outcomes $Y_{iM}(x)$. We also assume independence between X_{iM} and the sampling indicators, W_{iM}.

Assumption 8. (Assignment Mechanism) *The assignments $X_{1M}, \ldots, X_{MM}$ are independent, but not (necessarily) identically distributed, or inid.*

Because of the independence assumption we can apply laws of large numbers and central

limit theorems for *inid* (independent but not identically distributed) sequences. For the latter we rely on sufficient conditions for the Liapunov Central Limit Theorem.

To facilitate the asymptotic analysis we assume that the fourth moments of the triple (Y_{iM}, Z_{iM}, X_{iM}) are finite and uniformly bounded. We could relax this assumption at the cost of complicating the proofs. If we assume the sampling frequency ρ_M is bounded below by $\rho > 0$ we can get by with something less than uniformly bounded fourth moments, but here we want to include $\rho_M \to 0$ as a special case (leading to the EHW results) and keep the proofs transparent.

Assumption 9. (MOMENTS) *For all M the expected value $\mu_{k,l,m,M} = \mathbb{E}_{\boldsymbol{X}}[Y_{iM}^k \cdot X_{iM}^l \cdot Z_{iM}^m]$ is bounded by a common constant C for all nonnegative integers k, l, m such that $k + l + m \leq 4$.*

For convenience, we assume that the population moment matrices converge to fixed values. This is a technical simplification that could be relaxed, but relaxing it offers little in terms of substance. We also make a full rank assumption.

Assumption 10. (CONVERGENCE OF MOMENTS) *The sequences $\boldsymbol{Y}_M$, $\boldsymbol{Z}_M$ and $\boldsymbol{X}_M$ satisfy*

$$\Omega_M^{\text{pop}} = \mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{M}\sum_{i=1}^{M}\begin{pmatrix} Y_{iM} \\ Z_{iM} \\ X_{iM} \end{pmatrix}\begin{pmatrix} Y_{iM} \\ Z_{iM} \\ X_{iM} \end{pmatrix}'\right] \longrightarrow \Omega = \begin{pmatrix} \Omega_{YY} & \Omega_{YZ'} & \Omega_{YX'} \\ \Omega_{ZY} & \Omega_{ZZ'} & \Omega_{ZX'} \\ \Omega_{XY} & \Omega_{XZ'} & \Omega_{XX'} \end{pmatrix},$$

with Ω full rank.

For future reference define

$$\Gamma = \begin{pmatrix} \Omega_{XX} & \Omega_{XZ'} \\ \Omega_{ZX'} & \Omega_{ZZ'} \end{pmatrix}.$$

Given Assumption 10 we can define the limiting population estimands

$$\begin{pmatrix} \theta_\infty \\ \gamma_\infty \end{pmatrix} = \lim_{M\to\infty}\begin{pmatrix} \theta_M^{\text{causal}} \\ \gamma_M^{\text{causal}} \end{pmatrix} = \begin{pmatrix} \Omega_{XX} & \Omega_{XZ'} \\ \Omega_{ZX'} & \Omega_{ZZ'} \end{pmatrix}^{-1}\begin{pmatrix} \Omega_{XY} \\ \Omega_{ZY} \end{pmatrix} = \Gamma^{-1}\begin{pmatrix} \Omega_{XY} \\ \Omega_{ZY} \end{pmatrix}.$$

We maintain the random sampling assumption, Assumption 5. This implies that

$$\mathbb{E}_{\boldsymbol{W}}\left[\Omega_M^{\text{sample}}\middle| N, N > 0\right] = \Omega_M^{\text{pop}}.$$

In the proofs of the main results, we combine Assumptions 5 and 8, and use the fact that for all population sizes M, $\{(X_{iM}, W_{iM}) : i = 1, ..., M\}$ is an *inid* sequence where W_{iM} and X_{iM} are independent for all $i = 1, ..., M$, and all populations. We also maintain Assumption 3 concerning

the sampling rate, which guarantees that as the population size increases, the sample size N also tends to infinity. Allowing ρ_M to converge to zero allows for the possibility that the sample size is a neglible fraction of the population size: $\rho_M = \mathbb{E}[N]/M \to 0$, so the EHW results are included as a special case of our general results. Technically we should write N_M as the sample size but we drop the M subscript for notational convenience.

3.5 The General Case

First we state a result regarding the common limiting values of the least squares estimators and the causal and descriptive estimands:

Lemma 7. *Suppose Assumptions 3, 5 and 8-10 hold. Then* (i)

$$\begin{pmatrix} \hat{\theta}_{\text{ols}} - \theta_\infty \\ \hat{\gamma}_{\text{ols}} - \gamma_\infty \end{pmatrix} \xrightarrow{p} 0,$$

(ii)

$$\begin{pmatrix} \theta_M^{\text{descr}} - \theta_\infty \\ \gamma_M^{\text{descr}} - \gamma_\infty \end{pmatrix} \xrightarrow{p} 0,$$

and (iii)

$$\begin{pmatrix} \theta_M^{\text{causal}} - \theta_\infty \\ \gamma_M^{\text{causal}} - \gamma_\infty \end{pmatrix} \longrightarrow 0.$$

This result follows fairly directly from the assumptions about the moments and the sequence of populations, although allowing for the limiting case $\rho_M \to 0$ requires a little care in showing consistency of the least squares estimators. Note that part (iii) is about deterministic convergence and follows directly from Assumption 10 and the definition of the causal parameters.

Next we study the limiting distribution of the least squares estimator. The key component is the stochastic behavior of the normalized sample average of the product of the residuals and the covariates,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{M} W_{iM} \cdot \begin{pmatrix} X_{iM} \cdot \varepsilon_{iM} \\ Z_{iM} \cdot \varepsilon_{iM} \end{pmatrix}. \tag{3.1}$$

In our approach this normalized sum of independent but non-identically distributed terms has expectation zero – something we verify below – even though each of the separate terms $X_{iM} \cdot \varepsilon_{iM}$ and $Z_{iM} \cdot \varepsilon_{iM}$ may have non-zero expectations. To conclude that (3.1) has a limiting normal

distribution we must apply a central limit theorem for independent double arrays. Here we use the Liapunov central limit theorem as stated in Davidson (1994, Theorem 23.11).

Define the limits of the population quantities

$$\Delta_V = \lim_{M\to\infty} \mathbb{V}_{\boldsymbol{X}} \left(\frac{1}{\sqrt{M}} \sum_{i=1}^{M} \begin{pmatrix} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM} \end{pmatrix} \right), \tag{3.2}$$

$$\Delta_{\text{ehw}} = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{\boldsymbol{X}} \left[\begin{pmatrix} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM} \end{pmatrix} \begin{pmatrix} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM} \end{pmatrix}' \right] \tag{3.3}$$

and their difference

$$\Delta_E = \Delta_{\text{ehw}} - \Delta_V = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \left[\mathbb{E}_{\boldsymbol{X}} \begin{pmatrix} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM} \end{pmatrix} \right] \left[\mathbb{E}_{\boldsymbol{X}} \begin{pmatrix} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM} \end{pmatrix} \right]'. \tag{3.4}$$

Lemma 8. *Suppose Assumptions 3, 5, and 8-10 hold. Then:*

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{M} W_{iM} \cdot \begin{pmatrix} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM} \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \rho\cdot\Delta_V + (1-\rho)\cdot\Delta_{\text{ehw}}). \tag{3.5}$$

The first part of the asymptotic variance, $\rho\cdot\Delta_V$, captures the variation due to random assignment of the treatment. This component vanishes if the sample is small relative to the population. The second part, $(1-\rho)\cdot\Delta_{\text{ehw}}$, captures the variation due to random sampling. This is equal to zero if we observe the entire population.

Now we present the first of the two main results of the paper, describing the properties of the least squares estimator viewed as an estimator of the causal estimand and, separately, viewed as an estimator of the descriptive estimand:

Theorem 1. *Suppose Assumptions 3, 5 and 8-10 hold. Then* (i)

$$\sqrt{N} \begin{pmatrix} \hat{\theta}_{\text{ols}} - \theta_M^{\text{causal}} \\ \hat{\gamma}_{\text{ols}} - \gamma_M^{\text{causal}} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Gamma^{-1} \left(\Delta_{\text{ehw}} - \rho\cdot\Delta_E \right) \Gamma^{-1} \right),$$

(ii)

$$\sqrt{N} \begin{pmatrix} \hat{\theta}_{\text{ols}} - \theta_M^{\text{descr}} \\ \hat{\gamma}_{\text{ols}} - \gamma_M^{\text{descr}} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, (1-\rho)\cdot\Gamma^{-1}\Delta_{\text{ehw}}\Gamma^{-1} \right),$$

and (iii)

$$\sqrt{N} \begin{pmatrix} \theta_M^{\text{descr}} - \theta_M^{\text{causal}} \\ \gamma_M^{\text{descr}} - \gamma_M^{\text{causal}} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \rho\cdot\Gamma^{-1}\Delta_V\Gamma^{-1} \right).$$

Proof: See Appendix.

The standard EHW case is the special case in this theorem corresponding to $\rho = 0$. For both the causal and the descriptive estimand the asymptotic variance in the case with $\rho = 0$ reduces to $\Gamma^{-1}\Delta_{\text{ehw}}\Gamma^{-1}$. Moreover, the difference between the two estimands, θ_M^{causal} and θ_M^{desc}, normalized by the sample size, vanishes in this case. If the sample size is non-neglible as a fraction of the population sizes, $\rho > 0$, the difference between the EHW variance and the finite population causal variance is positive semi-definite, with the difference equal to $\rho \cdot \Gamma^{-1}\Delta_E\Gamma^{-1}$. This shows that the conventional robust sampling variance $\Gamma^{-1}\Delta_{\text{ehw}}\Gamma^{-1}$ is appropriate if either the sample size is small relative to population size, or if the expected values of $X_{iM} \cdot \varepsilon_{iM}$ and $Z_{iM} \cdot \varepsilon_{iM}$ are close to zero for all i (and thus Δ_E vanishes).

3.6 The Variance when the Regression Function is Correctly Specified

In general the difference between the causal variance and the conventional robust EHW variance, normalized by the sample size, is $\rho \cdot \Gamma^{-1}\Delta_E\Gamma^{-1}$. Here we investigate when the component of this difference corresponding to the causal effect θ_M^{causal} is equal to zero. The difference in variances obviously vanishes if the sample size is small relative to the population size, $\rho \approx 0$, but there is another interesting case where only the difference between the two variances that corresponds to the estimator for θ_M^{causal} vanishes, without Δ_E being equal to zero. This case arises when the regression function, as a function of the potential cause X_{iM}, is correctly specified. It is important to be explicit here about what we mean by "correctly specified." In the conventional approach, with random sampling from a large population, the notion of a correct specification is defined by reference to this large population. In that setting the linear specification is correct if the population average of the outcome for each value of the covariates lies on a straight line. Here we define the notion in the finite population where it need not be the case that there are multiple units with the same values for the covariates X_{iM} and Z_{iM} so that the large population definition does not apply. We make two specific assumptions. First, and this takes account of the potential causes part of the specification, we restrict the values of the potential outcomes. Second, and this takes account of the attributes part of the specification, we restrict the distribution of the assignments X_{iM}.

Assumption 11. (Linearity of Potential Outcomes) *The potential outcomes satisfy*

$$Y_{iM}(x) = Y_{iM}(0) + x'\theta.$$

Assumption 11 is not enough to conclude that the least squares estimator consistently estimates the causal parameters θ. We must also restrict the way in which the causes, X_{iM}, depend on $\{(Z_{iM}, Y_{iM}(0) : i = 1, 2, ..., M\}$, in an exogeneity or unconfoundedness-type assumption. To this end, define the vector of slope coefficients from the population regression $Y_{iM}(0)$ on Z_{iM}, $i = 1, 2, ..., M$, as

$$\gamma_M = \left(\frac{1}{M} \sum_{i=1}^{M} Z_{iM} Z_{iM}' \right)^{-1} \left(\frac{1}{M} \sum_{i=1}^{M} Z_{iM} Y_{iM}(0) \right). \tag{3.6}$$

This vector γ_M is non-stochastic because it depends only on attributes and potential outcomes.

Assumption 12. (Orthgonality of Assignment) *For all M,*

$$\sum_{i=1}^{M} (Y_{iM}(0) - Z_{iM}'\gamma_M) \cdot \mathbb{E}_{\boldsymbol{X}}\Big[X_{iM}\Big] = 0.$$

This assumption requires the mean of X_{iM} to be orthogonal to the population residuals $Y_{iM}(0) - Z_{iM}'\gamma_M$, which measure the part of $Y_{iM}(0)$ not explained by Z_{iM}. A special case is that where the X_{iM} are independent and identically distributed (as, for example, in a completely randomized experiment), so that $\mathbb{E}[X_{iM}] = \mu_X$, in which case Assumption 12 holds as long as there is an intercept in Z_{iM} because by the definition of γ_M, it follows that $\sum_{i=1}^{M}(Y_{iM}(0) - Z_{iM}'\gamma_M) = 0$. More interesting is another special case where $\mathbb{E}_{\boldsymbol{X}}[X_{iM}]$ is a linear function of Z_{iM}, say $\mathbb{E}_{\boldsymbol{X}}[X_{iM}] = \Lambda_M Z_{iM}$, $i = 1, ..., M$, for some matrix Λ_M. It is easily seen that in that case Assumption 12 holds because, by definition of γ_M,

$$\sum_{i=1}^{M} Z_{iM} \left(Y_{iM}(0) - Z_{iM}'\gamma_M\right) = 0.$$

In general, Assumption 12 allows X_{iM} to be systematically related to Z_{iM}, and even related to $Y_{iM}(0)$, provided the expected value of X_{iM} is uncorrelated in the population with the residual from regressing $Y_{iM}(0)$ on Z_{iM}'. Notice that only the first moments of the X_{iM} are restricted; the rest of the distributions are unrestricted.

Definition 1. *The regression function*

$$Y_{iM} = X'_{iM}\theta + Z'_{iM}\gamma + \varepsilon_{iM},$$

is correctly specified if Assumptions 11 and 12 hold.

Now we can establish the relationship between the population estimand θ_M^{causal} and the slope of the potential outcome function.

Theorem 2. *Suppose Assumptions 8, 9, 11, and 12 hold. Then for all M,*

$$\begin{pmatrix} \theta_M^{\text{causal}} \\ \gamma_M^{\text{causal}} \end{pmatrix} = \begin{pmatrix} \theta \\ \gamma_M \end{pmatrix}$$

Given Assumptions 11 and 12 we can immediately apply the result from Theorem 1 with θ instead of θ_M^{causal}, and we also have a simple interpretation for γ_M^{causal}.

An implication of Assumptions 11 and 12 is that the population residual ε_{iM} is no longer stochastic:

$$\begin{aligned} \varepsilon_{iM} &= Y_i(0) + X'_{iM}\theta - X'_{iM}\theta_M^{\text{causal}} - Z'_{iM}\gamma_M^{\text{causal}} \\ &= Y_i(0) - Z'_{iM}\gamma_M, \end{aligned}$$

which does not involve the stochastic components $\boldsymbol{X}_M$ or $\boldsymbol{W}_M$. This leads to simplifications in the variance components. The Γ component of the variance remains unchanged, but under Assumptions 11 and 12, Δ_V simplifies, with only the top-left block different from zero:

$$\Delta_V = \begin{pmatrix} \lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} \varepsilon_{iM}^2 \cdot \mathbb{V}_{\boldsymbol{X}}\left(X_{iM}\right) & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.7}$$

In order to simplify the asymptotic variance of $\sqrt{N}\left(\hat{\theta}_{\text{ols}} - \theta\right)$ we add the linearity assumption mentioned above.

Assumption 13. (Linearity of the Treatment In Attributes) *For some $K \times J$ matrix Λ_M, and for $i = 1, \ldots, M$,*

$$\mathbb{E}_{\boldsymbol{X}}[X_{iM}] = \Lambda_M Z_{iM}.$$

Recall that this assumption implies Assumption 12, and so we know least squares consistently estimates θ, and it has a limiting normal distribution when scaled by $\sqrt{N}$. But with Assumption

13 we can say more. Namely, the usual EHW variance is asymptotically valid for $\hat{\theta}_{\text{ols}} - \theta_M^{\text{causal}}$ (but remains conservative for $\hat{\gamma}_{\text{ols}} - \gamma_M^{\text{causal}}$). Define

$$\dot{X}_{iM} = X_{iM} - \Lambda_M Z_{iM}$$

Because under Assumptions 11 and 12 the residual ε_{iM} is non-stochastic it follows that

$$\mathbb{E}_{\boldsymbol{X}}\left[\dot{X}_{iM} \cdot \varepsilon_{iM}\right] = \mathbb{E}_{\boldsymbol{X}}\left[\dot{X}_{iM}\right] \cdot \varepsilon_{iM} = \left(\mathbb{E}_{\boldsymbol{X}}\left[X_{iM}\right] - \Lambda_M Z_{iM}\right) \cdot \varepsilon_{iM} = 0,$$

by Assumption 13.

Now define

$$\Gamma_{\dot{X}} = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{X}}\left[\dot{X}_{iM}\dot{X}_{iM}'\right],$$

and

$$\Delta_{\text{ehw},\dot{X}} = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{X}}\left[\varepsilon_{iM}^2 \dot{X}_{iM}\dot{X}_{iM}'\right].$$

Theorem 3. *Suppose Assumptions 8–13 hold. Then*

$$\sqrt{N}\left(\hat{\theta}_{\text{ols}} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \Gamma_{\dot{X}}^{-1} \Delta_{\text{ehw},\dot{X}} \Gamma_{\dot{X}}^{-1}\right).$$

The key insight in this theorem is that the asymptotic variance of $\hat{\theta}_{\text{ols}}$ does not depend on the ratio of the sample to the population size, ρ. We also know from Theorem 1 that if ρ is close to zero the proposed variance agrees with the EHW variance. Therefore it follows that the usual EHW variance matrix is correct for $\hat{\theta}_{\text{ols}}$ under these assumptions, and it can be obtained, as in standard asymptotic theory for least squares, by partially out Z_{iM} from X_{iM} in the population. For this result it is *not* sufficient that the regression function is correctly specified (Assumptions 11 and 12); we have also assumed linearity of the potential cause in the attributes (Assumption 13). Nevertheless, no other features of the distribution of X_{iM} are restricted.

For the case with X_{iM} binary and no attributes beyond the intercept this result can be inferred directly from Neyman's results for randomized experiments. In that case the focus is on the constant treatment assumption, which is extended to the linearity in Assumption 11. In that binary-treatment randomized-experiment case without attributes Assumptions 12 and 13 hold trivially. Generally, if linearity holds and X_{iM} is completely randomized then the conclusions of Theorem 3 hold.

The asymptotic variance of $\hat{\gamma}_{\text{ols}}$, the least squares estimates of the coefficients on the attributes, still depends on the ratio of sample to population size, and the the conventional robust EHW variance estimates over-estimates the uncertainty in these estimates.

4 Estimating the Variance

Now let us turn to the problem of estimating the variance for our descriptive and causal estimands. This is a complicated issue. The variance in the conventional setting is easy to estimate. One can consistently estimate Γ as the average of the matrix of outer products over the sample:

$$\hat{\Gamma} = \frac{1}{N}\sum_{i=1}^{M} W_{iM} \cdot \begin{pmatrix} Z_{iM} \\ X_{iM} \end{pmatrix} \begin{pmatrix} Z_{iM} \\ X_{iM} \end{pmatrix}'.$$

Also Δ_{ehw} is easy to estimate. First we estimate the residuals

$$\hat{\varepsilon}_{iM} = Y_{iM} - X_{iM}'\hat{\theta}_{\text{ols}} - Z_{iM}'\hat{\gamma}_{\text{ols}},$$

and then we estimate Δ_{ehw} as:

$$\hat{\Delta}_{\text{ehw}} = \frac{1}{N}\sum_{i=1}^{M} W_{iM} \cdot \begin{pmatrix} X_{iM}\cdot\hat{\varepsilon}_{iM} - \overline{X_{iM}\cdot\hat{\varepsilon}_{iM}} \\ Z_{iM}\cdot\hat{\varepsilon}_{iM} - \overline{Z_{iM}\cdot\hat{\varepsilon}_{iM}} \end{pmatrix} \begin{pmatrix} X_{iM}\cdot\hat{\varepsilon}_{iM} - \overline{X_{iM}\cdot\hat{\varepsilon}_{iM}} \\ Z_{iM}\cdot\hat{\varepsilon}_{iM} - \overline{Z_{iM}\cdot\hat{\varepsilon}_{iM}} \end{pmatrix}', \tag{4.1}$$

where

$$\overline{X_{iM}\cdot\hat{\varepsilon}_{iM}} = \frac{1}{N}\sum_{i=1}^{M} W_{iM}\cdot X_{iM}\cdot\hat{\varepsilon}_{iM}, \qquad \text{and} \quad \overline{Z_{iM}\cdot\hat{\varepsilon}_{iM}} = \frac{1}{N}\sum_{i=1}^{M} W_{iM}\cdot Z_{iM}\cdot\hat{\varepsilon}_{iM}.$$

In this case we do not need to subtract the averages, which in fact will be equal to zero, but this form is useful for subsequent variance estimators. The variance is then estimated as

$$\hat{V}_{\text{ehw}} = \hat{\Gamma}^{-1}\hat{\Delta}_{\text{ehw}}\Gamma^{-1}. \tag{4.2}$$

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

If we are interested in the descriptive estimand it is straightforward to modify the variance estimator. We simply multiply the 'ehww variance estimator by one minus the ratio of the sample size over the population size.

It is more challenging to estimate the variance of $\hat{\theta}_{\text{ols}} - \theta_M^{\text{causal}}$. The difficult is in estimating Δ_V (or, equivalently, $\Delta_E = \Delta_{\text{ehw}} - \Delta_V$). 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 2.3. In that case there are three terms in the expression for $\mathbb{V}^{\text{norm}}_{\text{causal}}$ presented in (2.3). The first two are straightforward to estimate, but the third one, $\sigma^2(\text{low}, \text{high})$ cannot be estimated consistently because we do not observe both potential outcomes for the same units. In that case researchers often use the conservative estimator based on ignoring that term. Here we can do the same. Because

$$\mathbb{V}_{\boldsymbol{X}}\left(\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\left(\begin{array}{c} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM}\end{array}\right)\right) \leq \mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{M}\sum_{i=1}^{M}\left(\begin{array}{c} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM}\end{array}\right)\left(\begin{array}{c} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM}\end{array}\right)'\right],$$

it follows that

$$\begin{aligned}\Delta_V &= \lim_{M\to\infty}\mathbb{V}_{\boldsymbol{X}}\left(\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\left(\begin{array}{c} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM}\end{array}\right)\right) \\ &\leq \lim_{M\to\infty}\mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{M}\sum_{i=1}^{M}\left(\begin{array}{c} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM}\end{array}\right)\left(\begin{array}{c} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM}\end{array}\right)'\right] = \Delta_{\text{ehw}},\end{aligned}$$

and we can use the estimator in (4.2) as the basis for a conservative estimator for the variance, $\hat{\Gamma}^{-1}\hat{\Delta}_{\text{ehw}}\hat{\Gamma}^{-1}$. However, we can do better. Instead of using the average of the outer product, $\hat{\Delta}_{\text{ehw}}$, as an upwardly biased estimator for Δ_V, we can remove part of the expected value.

Suppose we split the population into S strata, on the basis of the values of the non-stochastic variables $Y_{iM}(x)$ and Z_{iM}. Let $S_{iM} \in \{1,\ldots,S\}$, for $i = 1,\ldots,M$, $M = 1,2,\ldots$, be the indicator for the subpopulations, and let M_s be the stratum-specific population size. Then, by independence of the X_{iM} (and thus independence of the ε_{iM}), it follows that

$$\mathbb{V}_{\boldsymbol{X}}\left(\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\left(\begin{array}{c} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM}\end{array}\right)\right) = \sum_{s=1}^{S}\frac{M_s}{M}\cdot\mathbb{V}_{\boldsymbol{X}}\left(\frac{1}{\sqrt{M}_s}\sum_{i:S_{iM}=s}\left(\begin{array}{c} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM}\end{array}\right)\right).$$

Now we can obtain a conservative estimator of Δ_V by averaging within-stratum estimates after taking out with within-stratum averages:

$$\hat{\Delta}_{\text{strat}} = \sum_{s=1}^{S}\frac{M_s}{M}\cdot\hat{\Delta}_{\text{ehw},s},$$

where

$$\hat{\Delta}_{\text{ehw},s} = \hat{\mathbb{V}}_{\boldsymbol{X}}\left(\frac{1}{\sqrt{M}_s}\sum_{i:S_{iM}=s}\left(\begin{array}{c} X_{iM}\cdot\varepsilon_{iM} \\ Z_{iM}\cdot\varepsilon_{iM}\end{array}\right)\right) =$$

$$\frac{1}{M_s}\sum_{i:S_i=s}\left(\begin{pmatrix} X_{iM}\cdot\hat{\varepsilon}_{iM}-\overline{X_{iM}\cdot\hat{\varepsilon}_{iM}}_s \\ Z_{iM}\cdot\hat{\varepsilon}_{iM}-\overline{Z_{iM}\cdot\hat{\varepsilon}_{iM}}_s \end{pmatrix}\begin{pmatrix} X_{iM}\cdot\hat{\varepsilon}_{iM}-\overline{X_{iM}\cdot\hat{\varepsilon}_{iM}}_s \\ Z_{iM}\cdot\hat{\varepsilon}_{iM}-\overline{Z_{iM}\cdot\hat{\varepsilon}_{iM}}_s \end{pmatrix}'\right),$$

with

$$\overline{X_{iM}\cdot\hat{\varepsilon}_{iM}}_s=\frac{1}{M_s}\sum_{i:S_{iM}=s}X_{iM}\cdot\hat{\varepsilon}_{iM},\qquad\text{and}\quad\overline{Z_{iM}\cdot\hat{\varepsilon}_{iM}}_s=\frac{1}{M_s}\sum_{i:S_{iM}=s}Z_{iM}\cdot\hat{\varepsilon}_{iM}.$$

Assumption 14. *For $s=1,\ldots,S$, $M_s/M\to\delta_s>0$.*

Lemma 9. *Suppose Assumptions 8-14 hold. Then*

$$\Delta_V\leq\Delta_{\text{strat}}\leq\Delta_{\text{ehw}},$$

where

$$\Delta_{\text{strat}}=\text{plim}(\hat{\Delta}_{\text{strat}}).$$

The proposed estimator for the normalized variance is then

$$\hat{V}_{\text{strat}}=\hat{\Gamma}^{-1}\hat{\Delta}_{\text{strat}}\hat{\Gamma}^{-1}. \tag{4.3}$$

A natural way to define the strata is in terms of values of the attributes Z_{iM}. If the attributes are discrete we can simply stratify by their exact values. If the Z_{iM} take on many values we can partition the attribute space into a finite number of subspaces.

If one is willing to make the additional assumption that the potential outcome function is correctly specified, then we can make additional progress in estimating Δ_V. In that case only the top-left block of the Δ_V matrix differs from zero, as we discussed in Section 3.6. In addition, this assumption implies that the residuals ε_{iM} are non-stochastic, and so they can be used in partitioning the population.

5 Simulations

Here we present some evidence on the difference between the conventional ehw variance and the variance for causal effects. We focus on the case with a single binary cause $X_i\in\{-1,1\}$ and a single binary attribute $Z_i\in\{-1,1\}$. The potential outcome function has the form

$$Y_i(x)=Y_i(0)+(\tau_0+\tau_1\cdot Z_i+\tau_2\cdot\eta_i)\cdot x,$$

where τ_1 and τ_2 are parameters we vary across simulations. The $\eta_i \in \{-1, 1\}$ is an unobserved source of heterogeneity in the treatment effect. If $\tau_1 = \tau_2 = 0$, the regression function is correctly specified and because the attribute is binary the linearity condition is also satisfied, and thus the conventional ehw variance will be valid. If either τ_1 or τ_2 differs from zero, the conventional variance estimator will be over-estimating the variance. If τ_1 is different from zero the variance estimator based on partitioning the sample by values of Z_i will be an improvement over the conventional variance estimator.

In the population the Z_i and η_i are uncorrelated, and satisfy

$$\frac{1}{M}\sum_{i=1}^{M} \mathbf{1}_{\eta_i=-1} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}_{\eta_i=1} = 1/2, \qquad \text{and} \quad \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}_{Z_i=-1} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}_{Z_i=1} = 1/2.$$

Moreover,

$$\frac{1}{M}\sum_{i=1}^{M} Y_i(0) = 0, \qquad \frac{1}{M}\sum_{i=1}^{M} Y_i^2(0) = 1,$$

$$\frac{1}{M}\sum_{i=1}^{M} \eta_i \cdot Y_i(0) = \frac{1}{M}\sum_{i=1}^{M} Z_i \cdot Y_i(0) = \frac{1}{M}\sum_{i=1}^{M} Z_i \cdot \eta_i = 0,$$

and

$$\text{pr}(X_i = 1) = \text{pr}(X_i = -1) = 1/2.$$

The $Y_i(0)$ were generated by first drawing ν_i from a normal distribution with mean zero and variance equal to one, and then calculating $Y_i(0)$ as the residual from regressing ν_i on η_i, Z_i, and $\eta_i \cdot Z_i$.

We estimate a linear regression model using least squares:

$$Y_i = \gamma_0 + \gamma_1 \cdot Z_i + \theta \cdot X_i + \varepsilon_i.$$

We focus on the properties of the least squares estimator for θ.

We consider nine designs. In these designs we consider three sets of values for the pair (τ_1, τ_2), namely $(\tau_1 = 0, \tau_2 = 0)$, $(\tau_1 = 0, \tau_2 = 10)$, and $(\tau_1 = 10, \tau_2 = 0)$. In all cases $\tau_0 = 0$. The expected sample size $\rho \cdot M$ is in all cases equal to 1,000, but the value of ρ takes on different values, $\rho \in \{0.01, 0.5, 1\}$, so the population size is $M = 100,000$, $M = 2,000$, or $M = 1,000$ in the three different designs. In each of the nine designs we start by constructing a population,

as described above. Given the population we then repeatedly draw samples in the following two steps. We first randomly assign the covariates according to the binomial distribution with probability 1/2 for the two values $x = -1, 1$. Finally, we randomly sample units from this population, where each unit is sampled with probability ρ. For each unit in the sample we observe the triple (Y_i, X_i, Z_i).

Table 1: Simulation Results, $M \cdot \rho = 1000$

	$\tau_1 = 0, \quad \tau_2 = 0$			$\tau_1 = 0, \quad \tau_2 = 10$			$\tau_1 = 10, \quad \tau_2 = 0$		
	$\rho = .01$	$\rho = .5$	$\rho = 1$	$\rho = .01$	$\rho = .5$	$\rho = 1$	$\rho = .01$	$\rho = .5$	$\rho = 1$
$\text{std}(\theta^{\text{causal}})$	0.032	0.032	0.030	0.32	0.23	0.037	0.32	0.23	0.035
$\text{std}(\theta^{\text{desc}})$	0.031	0.023	0.000	0.32	0.23	0.000	0.32	0.22	0.000
se_{ehw}	0.032	0.032	0.032	0.32	0.32	0.318	0.32	0.32	0.318
$\text{se}_{\text{causal}}$	0.032	0.032	0.032	0.32	0.23	0.032	0.32	0.23	0.032
se_{desc}	0.032	0.022	0.000	0.32	0.22	0.000	0.32	0.22	0.000
$\widehat{\text{se}}_{\text{strat}}$	0.032	0.032	0.030	0.32	0.23	0.041	0.32	0.32	0.318

The results from the simulations are presented in Table 1. In the first two rows we present for each of the nine designs the standard deviation of the least squares estimator $\hat{\theta}_{\text{ols}}$ as an estimator for θ^{causal} and as an estimator for θ^{desc}. If $\rho = 0.01$ the two standard deviations are very similar, irrespective of the values of τ_1 and τ_2. If $\rho = 1$, the standard deviation of $\hat{\theta}_{\text{ols}} - \theta^{\text{desc}}$ is zero, whereas the standard deviation of $\hat{\theta}_{\text{ols}} - \theta^{\text{causal}}$ remains positive. If $\rho = 0.5$, the ratio of the variances depends on the other parameters of the design. The next three rows present the results of analytic calculations for the three variances, first the conventional ehw variance, then the variance for the causal estimand and finally the variance for the descriptive estimand. The latter two closely match the standard deviation of the estimator over the repeated samples, confirming that the theoretical calculations provide guidance for the sample sizes considered

here. Finally, in the last row we partition the sample by the values of Z_i, and use the variance estimator in (4.3). We see that in the case with $\tau_1 = \tau_2 = 0$ the properties of the proposed variance estimator are very similar to those of the conventional ehw estimator. If $\tau_1 > 0$ and $\tau_2 = 0$, so the variation in the coefficient on X_i is associated with the observed attribute Z_i, then the proposed variance estimator outperforms the conventional EHW estimator. If $\tau_1 = 0$ and $\tau_2 > 0$, so the variation in the coefficient on X_i is associated with the unobserved attributes, then the performance of the proposed variance estimator is similar to that of the conventional EHW estimator.

6 Inference for Alternative Questions

This paper 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 the year 2013. In words, we analyze inference for estimands of parameters that answer the following 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 2015?" Arguably in most empirical analyses that are intended to inform policy the object of interest depends on future, not just past, outcomes. This creates substantial problems for inference. Here we discuss some of the complications, but our main point is 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 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.) Given such a population the researcher may wish to estimate, say the difference in average 2015 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 2013. A natural estimator for the difference in average values for Northern and Southern countries in 2015 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 2015.

In order to construct confidence intervals for the difference in 2015, the researcher must make some assumptions about how country outcomes will vary from year to year. An extreme assumption is that outcomes in 2015 and 2013 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 2015 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. Moreoever, 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. In the absence of panel data the results are likely to be sensitive to such assumptions. We leave this direction for future work.

7 Conclusion

In this paper we study the interpretation of standard errors in regression analysis when the assumption that the sample is a random sample from a large population of interest is not attractive. The conventional robust standard errors justified by this 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 under this approach.

In the current paper 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 paper is only a first step in a broader research program. The concerns we have raised in this paper arise in many other settings and for other kinds of hypotheses, and the implications would need to be worked out for those settings. Section 6 suggests some directions we think are particularly natural to consider.

References

Abadie, A., A. Diamond, and J. Hainmueller, (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, Vol. 105(490), 493-505.

Abadie, A., G. Imbens, and F. Zheng, (2012), "Robust Inference for Misspecified Models Conditional on Covariates," NBER Working Paper.

Angrist, J., and S. Pischke, (2009), *Mostly Harmless Econometrics*, Princeton University Press, Princeton, NJ.

Barrios, T., R. Diamond, G. Imbens, and M. Kolesar, (2012), "Clustering, Spatial Correlations, and Randomization Inference," *Journal of the American Statistical Association*, Vol. 107(498): 578-591.

Bertrand, M., E. Duflo, and S. Mullainathan, (2004), "How Much Should We Trust Difference-In-Differences Estimates," *Quarterly Journal of Economics*, Vol. (119): 249-275.

Cattaneo, M., B. Frandsen, and R. Titiunik, (2013), "Randomization Inference in the Regression Discontinuity Design: An Application to the Study of Party Advantages in the U.S. Senate," Unpublished Working Paper.

Chow, G., (1984), "Maximum-likelihood estimation of misspecified models," *Economic Modelling*, Vol. 1(2): 134-138.

Cochran, W. (1969), "The Use of Covariance in Observational Studies," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 18(3): 270–275.

Cochran, W., (1977), *Sampling Techniques.* Wiley: New York.

Davidson, J., (1994), *Stochastic Limit Theory: An Introduction for Econometricicans*, Oxford University Press.

Deaton, A., (1997), *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy.* World Bank Publications.

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

Eicker, F., (1967), "Limit Theorems for Regression with Unequal and Dependent Errors," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 59-82, University of California Press, Berkeley.

Fisher, R. A., (1935), *The Design of Experiments*, 1st ed, Oliver and Boyd, London.

Frandsen, B., (2012), "Exact inference for a weak instrument, a small sample, or extreme quantiles," Unpublished Working Paper.

Freedman, D., (2008a), "On Regression Adjustments in Experiments with Several Treatments," *The Annals of Applied Statistics*, Vol. 2(1): 176–196.

Freedman, D., (2008b), "On Regression Adjustments to Experimental Data," *Advances in Applied Mathematics*, Vol. 40: 181–193.

Gelman, A., and J. Hill, (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press

Hayashi, F., (2000), *Econometrics*, Princeton University Press.

Holland, P., (1986), "Statistics and Causal Inference," (with discussion), *Journal of the American Statistical Association*, 81, 945-970.

Huber, P., (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 221-233, University of California Press, Berkeley.

Imbens, G., and P. Rosenbaum, (2005), "Robust, accurate confidence intervals with a weak instrument: quarter of birth and education," *Journal of the Royal Statistical Society, Series A (Theoretical Statistics)*, 168: 109-126.

Imbens, G.W., and J.M. Wooldridge, (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature* 47 (1), 5-86.

Kish, L., (1995), *Survey Sampling*, Wiley.

Lin, W., (2013), "Agnostic Notes on Regression Adjustments for Experimental Data: Reexamining Freedman's Critique," *The Annals of Applied Statistics*, Vol. 7:(1): 295–318.

Neyman, J., (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," translated in *Statistical Science* (with discussion), Vol 5, No 4, 465–480, 1990.

Rosenbaum, P., (1995), *Observational Studies*. Springer Verlag: New York.

Rosenbaum, P., (2002), "Covariance Adjustment in Randomized Experiments and Observational Studies," *Statistical Science*, Vol. 17:(3): 286–304.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

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

Schochet, P., (2010), "Is Regression Adjustment Supported by the Neyman Model for Causal Inference?" *Journal of Statistical Planning and Inference*, Vol. 140: 246–259.

White, H., (1980a), "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, Vol. 21(1):149-170.

White, H. (1980b), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

White, H., (1982), "Maximum likelihood estimation of misspecified models," *Econometrica*, Vol 50(1): l-25.

Appendix A: Proofs

Proof of Lemma 1: Conditional on $N = \sum_{i=1}^{M} W_i$ the vector $\boldsymbol{W}$ has a multinomial distribution with

$$\begin{aligned} \text{pr}(\boldsymbol{W} = \boldsymbol{w}|N) &= \binom{M}{N}^{-1}, \text{ for all } \boldsymbol{w} \text{ with } \sum_{j=1}^{N} w_j = N \\ &= 0, \text{ otherwise.} \end{aligned}$$

The expected value, variance and covariance of individual elements of W are given by

$$\begin{aligned} \mathbb{E}[W_j|N] &= \frac{N}{M} \\ \mathbb{V}(W_j|N) &= \left(\frac{N}{M}\right) \cdot \left(1 - \frac{N}{M}\right) = \frac{N \cdot (M-N)}{M^2} \\ \mathbb{C}(W_j, W_h|N) &= -\frac{N \cdot (M-N)}{M^2 \cdot (M-1)} \end{aligned}$$

Now consider the sample average

$$\hat{\mu}_M = \frac{1}{N} \sum_{j=1}^{M} W_j \cdot Y_j.$$

For notational simplicity we leave conditioning on $N > 0$ implicit. Then

$$\mathbb{E}[\hat{\mu}_M|N] = \frac{1}{N} \sum_{j=1}^{M} \mathbb{E}[W_j|N] \cdot Y_j = \frac{1}{N} \sum_{j=1}^{M} \left(\frac{N}{M}\right) \cdot Y_j = \frac{1}{M} \sum_{j=1}^{M} Y_j = \mu_M.$$

The sampling variance of $\hat{\mu}_M$ is can be obtained by writing $\hat{\mu}_M = \boldsymbol{W}'\boldsymbol{Y}/N$ so that

$$\mathbb{V}(\hat{\mu}_M|N) = \frac{1}{N^2} \boldsymbol{Y}' \mathbb{V}(\boldsymbol{W}|N) \boldsymbol{Y}.$$

From the conditional second moments of $\boldsymbol{W}$ it follows that

$$\boldsymbol{Y}' \mathbb{V}(\boldsymbol{W}|N) \boldsymbol{Y} = \left[\frac{N \cdot (M-N)}{M^2 \cdot (M-1)}\right] \boldsymbol{Y}' \begin{pmatrix} M-1 & -1 & -1 & & -1 \\ -1 & M-1 & & -1 & -1 \\ -1 & & & & \\ & & -1 & M-1 & -1 \\ -1 & & -1 & -1 & M-1 \end{pmatrix} \boldsymbol{Y}$$

Straightforward algebra shows that

$$\boldsymbol{Y}' \begin{pmatrix} M-1 & -1 & -1 & & -1 \\ -1 & M-1 & & -1 & -1 \\ -1 & & & & \\ & & -1 & M-1 & -1 \\ -1 & & -1 & -1 & M-1 \end{pmatrix} \boldsymbol{Y} = M \sum_{j=1}^{M} Y_j^2 - \left(\sum_{j=1}^{M} Y_j\right)^2 = M \cdot \sum_{i=1}^{M} (Y_j - \mu_M)^2,$$

so that

$$\boldsymbol{Y}'\mathbb{V}(\boldsymbol{W}|N)\boldsymbol{Y} = \frac{N\cdot(M-N)}{M}\left[\frac{1}{M-1}\sum_{j=1}^{M}(Y_j-\mu_M)^2\right] = \frac{N(M-N)}{M}\sigma_M^2.$$

Therefore,

$$\mathbb{V}(\hat{\mu}_M|N, N>0) = \frac{1}{N^2}\left[\frac{N\cdot(M-N)}{M}\cdot\sigma_M^2\right] = \frac{\sigma_M^2}{N}\cdot\left(1-\frac{N}{M}\right).$$

□

Note that this result generalizes to any set of constants $\{c_j\}_{j=1,\ldots,M}$, so that,

$$\mathbb{V}(\boldsymbol{W}'\boldsymbol{c}|N) = \boldsymbol{c}'\mathbb{V}(\boldsymbol{W}|N)\boldsymbol{c} = \frac{N\cdot(M-N)}{M\cdot(M-1)}\cdot\sum_{j=1}^{M}(c_j-\overline{c}_M)^2,$$

where $\overline{c}_M = \sum_{i=1}^{M} c_i/M$.

Before proving Lemma 2, we state a useful result.

Lemma A.1. *Suppose Assumptions 1 and 3 hold. Then:*

$$\frac{N}{M\cdot\rho_M} \xrightarrow{p} 1 \quad \textit{and} \quad \frac{M\cdot\rho_M}{N} \xrightarrow{p} 1 \quad \textit{as } M\longrightarrow\infty.$$

Proof: Under Assumption 1, for any positive integer M, $N \sim Binomial(M, \rho_M)$, which implies $\mathbb{E}_{\boldsymbol{W}}[N] = M\cdot\rho_M$ and $\mathbb{V}_{\boldsymbol{W}}(N) = M\cdot\rho_M\cdot(1-\rho_M)$. Therefore,

$$\mathbb{E}_{\boldsymbol{W}}\left[\frac{N}{M\cdot\rho_M}\right] = 1 \qquad \text{and} \quad \mathbb{V}_{\boldsymbol{W}}\left(\frac{N}{M\cdot\rho_M}\right) = \frac{M\cdot\rho_M\cdot(1-\rho_M)}{M^2\cdot\rho_M^2} = \frac{(1-\rho_M)}{M\cdot\rho_M},$$

which converges to zero by Assumption 3. Therefore convergence in probability follows from convergence in mean square. The second part follows from Slutsky's Theorem because the reciprocal function is continuous at all nonzero values. □

Proof of Lemma 2: From Lemma 1,

$$\mathbb{V}_{\boldsymbol{W}}(\hat{\mu}_M|N, N>0) = \frac{\sigma_M^2}{N}\cdot\left(1-\frac{N}{M}\right),$$

and so

$$\mathbb{V}_{\boldsymbol{W}}(\hat{\mu}_M|N, N>0) - \frac{\sigma^2}{N} = \frac{\sigma_M^2-\sigma^2}{N} - \frac{\sigma_M^2}{M} = \frac{\sigma_M^2-\sigma^2}{\rho_M\cdot M} - \frac{\sigma_M^2}{M} - \frac{\sigma_M^2-\sigma^2}{\rho_M\cdot M}\cdot\left(1-\frac{\rho_M\cdot M}{N}\right).$$

Given Assumption 2, $\sigma_M^2 \to \sigma^2$ as $M\to\infty$, and therefore $\{\sigma_M^2\}$ is bounded. It follows that $\mathbb{V}_{\boldsymbol{W}}(\hat{\mu}_M|N, N>0) - \sigma^2/N = O_p((\rho_M M)^{-1})$, finishing the proof of part (i).

The normalized variance is

$$\mathbb{V}_{\boldsymbol{W}}^{\text{norm}}(\hat{\mu}_M|N) = N \cdot \mathbb{V}_{\boldsymbol{W}}(\hat{\mu}_M|N) = \sigma_M^2 \left(1 - \frac{N}{M}\right)$$

By Assumption 2, $\sigma_M^2 \to \sigma^2$. By Assumption 1, $N \sim Binomial(M, \rho_M)$ and so $\mathbb{E}(N/M) = \rho_M$ and

$$\mathbb{V}\left(\frac{N}{M}\right) = \frac{\rho_M(1-\rho_M)}{M} \to 0,$$

which means that $N/M - \rho_M \xrightarrow{p} 0$. Along with Assumption 3 ($\rho_M \to \rho$) we get $\mathbb{V}_{\boldsymbol{W}}^{\text{norm}}(\hat{\mu}_M|N) \xrightarrow{p} \sigma^2(1-\rho)$.

Proof of Lemma 3: Assumption 1 ensures that the vector of sampling indicators over the two subpopulations, of size M_{coast} and $M_{\wedge}$ are independent. Further, conditional on N_{coast} and $N_{\wedge}$, they have the multinomial distribution described in the proof of Lemma 1. The result follows immediately because the covariance between the two sample means, conditional on $(N_{\text{coast}}, N_{\wedge})$ and $N_{\text{coast}} > 0$ and $N_{\wedge} > 0$, is zero. □

Proof of Lemma 4: Conditional on M_{low}, $M_{\text{high}} > 0$, write θ_M^{descr} as

$$\theta_M^{\text{descr}} = \frac{1}{M_{\text{high}}} \sum_{i=1}^{M} 1_{X_i=\text{high}} \cdot Y_i(\text{high}) - \frac{1}{M_{\text{low}}} \sum_{i=1}^{M} 1_{X_i=\text{low}} \cdot Y_i(\text{low})$$

Conditional on M_{high} (and therefore conditional on both M_{high} and M_{low}), $\mathbb{E}[1_{X_i=\text{high}}|M_{\text{high}}] = \text{pr}(X_i = \text{high}|M_{\text{high}}) = M_{\text{high}}/M$, and so

$$\mathbb{E}\left[\theta_M^{\text{descr}}|M_{\text{high}}, M_{\text{low}}\right] = \frac{1}{M_{\text{high}}} \sum_{i=1}^{M} \frac{M_{\text{high}}}{M} \cdot Y_i(\text{high}) - \frac{1}{M_{\text{low}}} \sum_{i=1}^{M} \frac{M_{\text{low}}}{M} \cdot Y_i(\text{low})$$

$$= \frac{1}{M} \sum_{i=1}^{M} \Big(Y_i(\text{high}) - Y_i(\text{low})\Big) = \theta_M^{\text{causal}}.$$

To compute the variance of θ_M^{descr}, write

$$\theta_M^{\text{descr}} = \sum_{i=1}^{M} 1_{X_i=\text{high}} \cdot \left(\frac{Y_i(\text{high})}{M_{\text{high}}} + \frac{Y_i(\text{low})}{M_{\text{low}}}\right) - \sum_{i=1}^{M} \frac{Y_i(\text{low})}{M_{\text{low}}}$$

Conditional on M_{low}, $M_{\text{high}} > 0$, the calculation is very similar to that in Lemma 1. In fact, take

$$c_i = \frac{Y_i(\text{high})}{M_{\text{high}}} + \frac{Y_i(\text{low})}{M_{\text{low}}},$$

and then Lemma 1 implies

$$\mathbb{V}\left(\sum_{i=1}^{M} X_i \left[\left(\frac{Y_i(\text{high})}{M_{\text{high}}}\right) + \left(\frac{Y_i(\text{low})}{M_{\text{low}}}\right)\right] \middle| M_{\text{low}}, M_{\text{high}}\right)$$

$$= \frac{M_{\text{low}}M_{\text{high}}}{M}\left[\frac{1}{M_{\text{high}}^2}\sigma^2(\text{high}) + \frac{1}{M_{\text{low}}^2}\sigma^2(\text{low})\right]$$

$$+\frac{2}{M(M-1)}\sum_{i=1}^{M}\Big(Y_i(\text{high}) - \bar{Y}(\text{high})\Big)\cdot\Big(Y_i(\text{low}) - \bar{Y}(\text{low})\Big).$$

Now

$$\sigma^2(\text{low},\text{high}) = \frac{1}{M-1}\sum_{i=1}^{M}[(Y_i(\text{high}) - \bar{Y}(\text{high})) - (Y_i(\text{low}) - \bar{Y}(\text{low}))]^2$$

$$= \sigma^2(\text{high}) + \sigma^2(\text{low}) - 2(M-1)^{-1}\sum_{i=1}^{M}[Y_i(\text{high}) - \bar{Y}(\text{high})][Y_i(\text{low}) - \bar{Y}(\text{low})].$$

or

$$2(M-1)^{-1}\sum_{i=1}^{M}[Y_i(\text{high}) - \bar{Y}(\text{high})][Y_i(\text{low}) - \bar{Y}(\text{low})] = \sigma^2(\text{high}) + \sigma^2(\text{low}) - \sigma^2(\text{low},\text{high}).$$

Substituting gives

$$\begin{aligned}
\mathbb{V}\left(\theta_M^{\text{descr}}|M_{\text{low}}, M_{\text{high}}\right) &= \frac{M_{\text{low}}M_{\text{high}}}{M}\left[\frac{1}{M_{\text{high}}^2}\sigma^2(\text{high}) + \frac{1}{M_{\text{low}}^2}\sigma^2(\text{low}) + \frac{[\sigma^2(\text{high}) + \sigma^2(\text{low}) - \sigma^2(\text{low},\text{high})]}{M_{\text{low}}M_{\text{high}}}\right] \\
&= \frac{M_{\text{low}}M_{\text{high}}}{M}\left[\frac{M}{M_{\text{low}}M_{\text{high}}^2}\sigma^2(\text{high}) + \frac{M}{M_{\text{high}}M_{\text{low}}^2}\sigma^2(\text{low}) - \frac{\sigma^2(\text{low},\text{high})}{M_{\text{low}}M_{\text{high}}}\right] \\
&= \frac{\sigma^2(\text{high})}{M_{\text{high}}} + \frac{\sigma^2(\text{low})}{M_{\text{low}}} - \frac{\sigma^2(\text{low},\text{high})}{M}.
\end{aligned}$$

□

Proof of Lemma 5: We prove parts (i) and (ii), as the other parts are similar (and (v) follows immediately). First, because $\boldsymbol{X}$ and $\boldsymbol{W}$ are independent, we have

$$\mathbb{D}(\boldsymbol{X}|\boldsymbol{W}, N_{\text{high}}, N_{\text{low}}) = \mathbb{D}(\boldsymbol{X}|N_{\text{high}}, N_{\text{low}})$$

and the distribution is multinomial with

$$\begin{aligned}
\mathbb{E}[1_{X_i=\text{high}}|N_{\text{high}}, N_{\text{low}}] &= \mathbb{E}[1_{X_i=\text{high}}|N_{\text{high}}, N) = N_{\text{high}}/N \\
\mathbb{V}(1_{X_i=\text{high}}|N_{\text{high}}, N) &= \frac{N_{\text{high}}N_{\text{low}}}{N^2} \\
\mathbb{C}(1_{X_i=\text{high}}, 1_{X_h=\text{high}}|N_{\text{high}}, N) &= -\frac{N_{\text{high}}N_{\text{low}}}{N^2(N-1)} \\
\mathbb{E}[1_{X_i=\text{high}}\cdot 1_{X_h=\text{high}}|N_{\text{high}}, N] &= \frac{N_{\text{high}}(N_{\text{high}}-1)}{N(N-1)}
\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}[1_{X_i=\text{high}} \cdot 1_{X_h=\text{low}}|N_{\text{high}}, N] &= \frac{N_{\text{high}}}{N} - \frac{N_{\text{high}}(N_{\text{high}}-1)}{N(N-1)} = \frac{(N-1)N_{\text{high}} - N_{\text{high}}^2 + N_{\text{high}}}{N(N-1)} \\ &= \frac{NN_{\text{high}} - N_{\text{high}}^2}{N(N-1)} = \frac{N_{\text{high}}(N-N_{\text{high}})}{N(N-1)} = \frac{N_{\text{high}}N_{\text{low}}}{N(N-1)}\end{aligned}$$

Now

$$\begin{aligned}\hat{\theta}_M &= N_{\text{high}}^{-1}\sum_{i=1}^{M} W_i 1_{X_i=\text{high}} Y_i(\text{high}) - N_{\text{low}}^{-1}\sum_{i=1}^{M} W_i 1 - X_i = \text{low} Y_i(\text{low}) \\ \mathbb{E}(\hat{\theta}_M|\boldsymbol{W}, N_{\text{high}}, N_{\text{low}}) &= \frac{1}{N_{\text{high}}}\sum_{i=1}^{M} W_i \text{pr}(X_i = \text{high}|N_{\text{high}}, N_{\text{low}}) Y_i(\text{high}) \\ &\qquad -\frac{1}{N_{\text{low}}}\sum_{i=1}^{M} W_i \text{pr}(X_i = \text{low}|N_{\text{high}}, N_{\text{low}})] Y_i(\text{low}) \\ &= N_{\text{high}}^{-1}\sum_{i=1}^{M} W_i (N_{\text{high}}/N) Y_i(\text{high}) - N_{\text{low}}^{-1}\sum_{i=1}^{M} W_i (N_{\text{low}}/N) Y_i(\text{low}) \\ &= N^{-1}\sum_{i=1}^{M} W_i [Y_i(\text{high}) - Y_i(\text{low})]\end{aligned}$$

and so

$$\begin{aligned}\mathbb{E}(\hat{\theta}_M|N_{\text{high}}, N_{\text{low}}) &= N^{-1}\mathbb{E}(W_i|N_{\text{high}}, N_{\text{low}})[Y_i(\text{high}) - Y_i(\text{low})] = N^{-1}\sum_{i=1}^{M}(N/M)[Y_i(\text{high}) - Y_i(\text{low})] \\ &= \mu_{M\text{high}} - \mu_{M\text{low}} = \theta_M^{\text{causal}},\end{aligned}$$

which proves part (i).
For part (ii) we find

$$\mathbb{V}(\hat{\theta}_M|N_{\text{high}}, N_{\text{low}}) = \mathbb{V}(\bar{Y}_{\text{high}}|N_{\text{high}}, N_{\text{low}}) + \mathbb{V}(\bar{Y}_{\text{low}}|N_{\text{high}}, N_{\text{low}}) - 2\mathbb{C}(\bar{Y}_{\text{high}}, \bar{Y}_{\text{low}}|N_{\text{high}}, N_{\text{low}}).$$

If we define $Z_i = W_i \cdot 1_{X_i=\text{high}}$ and $R_i = W_i \cdot 1_{X_i=\text{low}}$ then we can apply Lemma 4 to obtain the variances because

$$(Z_1, ..., Z_M)|(N_{\text{low}}, N_{\text{high}})$$

has a multinomial distribution with $\text{pr}(Z_i = 1|N_{\text{low}}, N_{\text{high}}) = N_{\text{high}}/M$ and $(R_1, ..., R_M)|(N_{\text{low}}, N_{\text{high}})$ has the distribution with $P(R_i = 1|N_{\text{low}}, N_{\text{high}}) = N_{\text{low}}/N$. Therefore,

$$\begin{aligned}\mathbb{V}(\bar{Y}_{\text{high}}|N_{\text{high}}, N_{\text{low}}) &= \frac{\sigma^2(\text{high})}{N_{\text{high}}}\left(1 - \frac{N_{\text{high}}}{M}\right) = \frac{\sigma^2(\text{high})}{N_{\text{high}}} - \frac{\sigma^2(\text{high})}{M} \\ \mathbb{V}(\bar{Y}_{\text{low}}|N_{\text{high}}, N_{\text{low}}) &= \frac{\sigma^2(\text{low})}{N_{\text{low}}} - \frac{\sigma^2(\text{low})}{M}\end{aligned}$$

and so

$$\mathbb{V}(\hat{\theta}|N_{\text{high}}, N_{\text{low}}) = \frac{\sigma^2(\text{high})}{N_{\text{high}}} + \frac{\sigma^2(\text{low})}{N_{\text{low}}} - \frac{\sigma^2(\text{high}) + \sigma^2(\text{low})}{M} - 2\mathbb{C}(\bar{Y}_{\text{high}}, \bar{Y}_{\text{low}}|N_{\text{high}}, N_{\text{low}})$$

We showed in 4 that

$$\begin{aligned}\sigma^2(\text{high})+\sigma^2(\text{low}) &= \sigma^2_{\text{low,high}}+\frac{2}{(M-1)}\sum_{i=1}^{M}[Y_i(\text{high})-\mu_{\text{high}}][Y_i(\text{low})-\mu_{\text{low}}]\\ &\equiv \sigma^2_{\text{low,high}}+2\eta_{\text{low,high}}\end{aligned}$$

where $\eta_{\text{low,high}}$ is the population covariance of $Y_i(\text{low})$ and $Y_i(\text{high})$. So

$$\mathbb{V}(\hat{\theta}|N_{\text{high}},N_{\text{low}})=\frac{\sigma^2(\text{high})}{N_{\text{high}}}+\frac{\sigma^2(\text{low})}{N_{\text{low}}}-\frac{\sigma^2_{\text{low,high}}}{M}-2\left[\frac{\eta_{low,\text{high}}}{M}+\mathbb{C}(\bar{Y}_{\text{high}},\bar{Y}_{\text{low}}|N_{\text{high}},N_{\text{low}})\right]$$

The proof is complete if we show

$$\mathbb{C}(\bar{Y}_{\text{high}},\bar{Y}_{\text{low}}|N_{\text{high}},N_{\text{low}})=-\frac{\eta_{\text{low,high}}}{M}$$

The usual algebra of covariances gives

$$\frac{\eta_{\text{low,high}}}{M}=\frac{1}{M(M-1)}\sum_{i=1}^{M}Y_i(\text{high})Y_i(\text{low})-\frac{\mu_{\text{high}}\mu_{\text{low}}}{(M-1)}$$

and so it suffices to show

$$\mathbb{E}(\bar{Y}_{\text{high}}\bar{Y}_{\text{low}}|N_{\text{high}},N_{\text{low}})-\mu_{\text{high}}\mu_{\text{low}}=\frac{\mu_{\text{high}}\mu_{\text{low}}}{(M-1)}-\frac{1}{M(M-1)}\sum_{i=1}^{M}Y_i(\text{high})Y_i(\text{low})$$

or

$$\begin{aligned}\mathbb{E}(\bar{Y}_{\text{high}}\bar{Y}_{\text{low}}|N_{\text{high}},N_{\text{low}}) &= \frac{M\mu_{\text{high}}\mu_{\text{low}}-M^{-1}\sum_{i=1}^{M}Y_i(\text{high})Y_i(\text{low})}{(M-1)}\\ &= \frac{\left(\sum_{i=1}^{M}Y_i(\text{high})\right)\left(\sum_{i=1}^{M}Y_i(\text{low})\right)-\left(\sum_{i=1}^{M}Y_i(\text{high})Y_i(\text{low})\right)}{M(M-1)}\\ &= \frac{\sum_{i=1}^{M}\sum_{h\neq i+1}^{M}Y_i(\text{high})Y_h(\text{low})}{M(M-1)}.\end{aligned}$$

To show this equivalance, write

$$\begin{aligned}\bar{Y}_{\text{high}}\bar{Y}_{\text{low}} &= \frac{1}{N_{\text{high}}N_{\text{low}}}\left(\sum_{i=1}^{M}W_i1_{X_i=\text{high}}Y_i(\text{high})\right)\left(\sum_{h=1}^{M}W_h1_{X_i=\text{low}}Y_h(\text{low})\right)\\ &= \frac{1}{N_{\text{high}}N_{\text{low}}}\left(\sum_{i=1}^{M}\sum_{h\neq i}^{M}W_i1_{X_i=\text{high}}Y_i(\text{high})W_h1_{X_h=\text{low}}Y_h(\text{low})\right)\end{aligned}$$

First condition on the sampling indicators $\boldsymbol{W}$ as well as $(N_{\text{high}}, N_{\text{low}})$:

$$\begin{aligned}
\mathbb{E}(\bar{Y}_{\text{high}}\bar{Y}_{\text{low}}|\boldsymbol{W}, N_{\text{high}}, N_{\text{low}}) &= \frac{1}{N_{\text{high}}N_{\text{low}}}\left(\sum_{i=1}^{M}\sum_{h\neq i}^{M} W_i W_h \text{pr}(X_i = \text{high}, X_h = \text{low}|\boldsymbol{W}, N_{\text{high}}, N_{\text{low}})Y_i(\text{high})Y_h(\text{low})\right. \\
&= \frac{1}{N_{\text{high}}N_{\text{low}}}\left(\sum_{i=1}^{M}\sum_{h\neq i}^{M} W_i W_h \text{pr}(X_i = \text{high}, X_h = \text{low}|N_{\text{high}}, N_{\text{low}})Y_i(\text{high})Y_h(\text{low})\right) \\
&= \frac{1}{N_{\text{high}}N_{\text{low}}}\left(\sum_{i=1}^{M}\sum_{h\neq i}^{M} W_i W_h [N_{\text{high}}N_{\text{low}}/N(N-1)]Y_i(\text{high})Y_h(\text{low})\right) = \\
&= \frac{1}{N(N-1)}\sum_{i=1}^{M}\sum_{h\neq i}^{M} W_i W_h Y_i(\text{high})Y_h(\text{low}).
\end{aligned}$$

Finally, use iterated expectations:

$$\begin{aligned}
E(\bar{Y}_{\text{high}}\bar{Y}_{\text{low}}|N_{\text{high}}, N_{\text{low}}) &= \frac{1}{N(N-1)}\sum_{i=1}^{M}\sum_{h\neq i}^{M} E(W_i W_h|N_{\text{high}}, N_{\text{low}})Y_i(\text{high})Y_h(\text{low}) \\
&= \frac{1}{N(N-1)}\sum_{i=1}^{M}\sum_{h\neq i}^{M} [N(N-1)/M(M-1)]Y_i(\text{high})Y_h(\text{low}) \\
&= \frac{1}{N(N-1)}\sum_{i=1}^{M}\sum_{h\neq i}^{M} [N(N-1)/M(M-1)]Y_i(\text{high})Y_h(\text{low}) \\
&= \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{h\neq i}^{M} Y_i(\text{high})Y_h(\text{low}),
\end{aligned}$$

which is what we needed to show. □

Proof of Lemma 6:

$$\mathbb{V}_{\text{causal}}^{\text{norm}} = \frac{\sigma_M^2(\text{low})}{N_{\text{low}}/N} + \frac{\sigma_M^2(\text{high})}{N_{\text{high}}/N} - \frac{N}{M}\cdot\sigma_M^2(\text{low}, \text{high}).$$

By Assumption 3 $N/M \to \rho$. By Assumptions 3 and 7 $N_{\text{low}}/N \to (1-q)$ and $N_{\text{high}}/N \to q$ By Assumption 6

$$\sigma_M^2(\text{low}) \longrightarrow \sigma^2(\text{low}), \qquad \sigma_M^2(\text{high}) \longrightarrow \sigma^2(\text{high}), \qquad \sigma_M^2(\text{low}, \text{high}) \longrightarrow \sigma^2(\text{low}, \text{high}).$$

Together these imply the two results in the lemma. □

It is useful to state a lemma that we use repeatedly in the asymptotic theory.

Lemma A.2. *For a sequence of random variables $\{U_{iM} : i = 1, ..., M\}$ assume that $\{(W_{iM}, U_{iM}) : i = 1, ..., M\}$ is independent but not (necessarily) identically distributed. Further, W_{iM} and U_{iM} are*

independent for all i=1,...,M. Assume that $\mathbb{E}(U_{iM}^2) < \infty$ *for* $i = 1, ..., M$ *and*

$$\begin{aligned} M^{-1}\sum_{i=1}^{M}\mathbb{E}(U_{iM}) &\rightarrow \mu_U \\ M^{-1}\sum_{i=1}^{M}\mathbb{E}(U_{iM}^2) &\rightarrow \kappa_U^2 \end{aligned}$$

Finally, assume that Assumptions 1 and 3 hold. Then

$$N^{-1}\sum_{i=1}^{M} W_{iM}U_{iM} - M^{-1}\sum_{i=1}^{M}\mathbb{E}(U_{iM}) \xrightarrow{p} 0.$$

Proof: Write the first average as

$$N^{-1}\sum_{i=1}^{M} W_{iM}U_{iM} = \left(\frac{M\rho_M}{N}\right) M^{-1}\sum_{i=1}^{M}\left(\frac{W_{iM}}{\rho_M}\right) U_{iM}.$$

As argued in the text, because $N \sim Binomial(M, \rho_M)$ and $M\rho_M \rightarrow \infty$ by Assumption 3, $(M\rho_M)/N \xrightarrow{p} 1$. Because we assume $M^{-1}\sum_{i=1}^{M}\mathbb{E}(U_{iM})$ converges, it is bounded, and so it suffices to show that

$$M^{-1}\sum_{i=1}^{M}\left(\frac{W_{iM}}{\rho_M}\right) U_{iM} - M^{-1}\sum_{i=1}^{M}\mathbb{E}(U_{iM}) \xrightarrow{p} 0$$

Now because W_{iM} is indepenent of U_{iM},

$$\mathbb{E}\left[M^{-1}\sum_{i=1}^{M}\left(\frac{W_{iM}}{\rho_M}\right) U_{iM}\right] = M^{-1}\sum_{i=1}^{M}\left(\frac{\mathbb{E}(W_{iM})}{\rho_M}\right)\mathbb{E}(U_{iM}) = M^{-1}\sum_{i=1}^{M}\mathbb{E}(U_{iM}),$$

and so the expected value of

$$M^{-1}\sum_{i=1}^{M}\left(\frac{W_{i}}{\rho_M}\right) U_{iM} - M^{-1}\sum_{i=1}^{M}\mathbb{E}(U_{iM})$$

is zero. Further, its variance exists by the second moment assumption, and by independence across i,

$$\begin{aligned} \mathbb{V}\left[M^{-1}\sum_{i=1}^{M}\left(\frac{W_{iM}}{\rho_M}\right) U_{iM}\right] &= M^{-2}\sum_{i=1}^{M}\frac{1}{\rho_M^2}\mathbb{V}(W_{iM}U_{iM}) = M^{-2}\sum_{i=1}^{M}\left\{\frac{1}{\rho_M^2}\mathbb{E}[(W_{iM}U_{iM})^2] - [\mathbb{E}(W_{iM}U_{iM})]^2\right\} \\ &= M^{-2}\sum_{i=1}^{M}\left\{\frac{1}{\rho_M^2}\rho_M\mathbb{E}(U_{iM}^2) - \rho_M^2[\mathbb{E}(U_{iM})]^2\right\} \leq M^{-2}\rho_M^{-1}\sum_{i=1}^{M}\mathbb{E}(U_{iM}^2) \\ &= \frac{1}{M\rho_M}\left[M^{-1}\sum_{i=1}^{M}\mathbb{E}(U_{iM}^2)\right]. \end{aligned}$$

By assumption, the term in brackets converges and by Assumption 3 $M\rho_M \rightarrow \infty$. We have shown mean square convergence and so convergence in probability follows. □

We can apply the previous lemma to the second moment matrix of the data. Define

$$\hat{\Omega}_M = \frac{1}{N}\sum_{i=1}^{M} W_{iM} \cdot \begin{pmatrix} Y_{iM}^2 & Y_{iM}X_{iM}' & Y_{iM}Z_i' \\ X_{iM}Y_{iM} & X_{iM}X_{iM}' & X_{iM}Z_{iM}' \\ Z_{iM}Y_{iM} & Z_{iM}X_{iM}' & Z_{iM}Z_{iM}' \end{pmatrix}.$$

Lemma A.3. *Suppose Assumptions 8–10 hold. Then:*

$$\hat{\Omega}_M - \Omega_M \xrightarrow{p} 0.$$

Proof: This follows from the previous lemma by letting U_{iM} be an element of the above matrix in the summand. The moment conditions are satisfied by Assumption 9 because fourth moments are assumed to be finite. □

Note that in combination with the assumption that $\lim_{M\to\infty}\Omega_M = \Omega$, Lemma A.3 implies that

$$\hat{\Omega}_M \xrightarrow{p} \Omega. \tag{A.1}$$

Proof of Lemma 7: The first claim follows in a straightforward manner from the assumptions and Lemma A.3 because the OLS estimators can be written as

$$\begin{pmatrix} \hat{\theta}_{\text{ols}} \\ \hat{\gamma}_{\text{ols}} \end{pmatrix} = \begin{pmatrix} \hat{\Omega}_{XX,M}^{\text{sample}} & \hat{\Omega}_{XZ',M}^{\text{sample}} \\ \hat{\Omega}_{ZX',M}^{\text{sample}} & \hat{\Omega}_{ZZ',M}^{\text{sample}} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\Omega}_{XY,M}^{\text{sample}} \\ \hat{\Omega}_{ZY,M}^{\text{sample}} \end{pmatrix}.$$

We know each element in the $\hat{\Omega}_M$ converges, and we assume its probability limit is positive definite. The result follows. The other claims are even easier to verify because they do not involve the sampling indicators W_{iM}. □

Next we prove a lemma that is useful for establishing asympotic normality.

Lemma A.4. *For a sequence of random variables $\{U_{iM} : i = 1, ..., M\}$ assume that $\{(W_{iM}, U_{iM}) : i = 1, ..., M\}$ is independent but not (necessarily) identically distributed. Further, W_{iM} and U_{iM} are independent for all i=1,...,M. Assume that for some $\delta > 0$ and $D < \infty$, $\mathbb{E}(|U_{iM}|^{2+\delta}) \leq D$ and $\mathbb{E}(|U_{iM}|) \leq D$, for $i = 1, ..., M$ and all M. Also,*

$$M^{-1}\sum_{i=1}^{M}\mathbb{E}[U_{iM}] = 0$$

and

$$\begin{aligned} \sigma_{U,M}^2 &= M^{-1}\sum_{i=1}^{M}\mathbb{V}(U_{iM}) \to \sigma_U^2 > 0 \\ \kappa_{U,M}^2 &= M^{-1}\sum_{i=1}^{M}[\mathbb{E}(U_{iM})]^2 \to \kappa_U^2. \end{aligned}$$

Finally, assume that Assumptions 1 and 3 hold. Then

$$N^{-1/2}\sum_{i=1}^{M} W_{iM}U_{iM} \xrightarrow{d} \mathcal{N}\left(0, [\sigma_U^2 + (1-\rho)\kappa_U^2]\right).$$

Proof: First, write

$$N^{-1/2}\sum_{i=1}^{M} W_{iM}U_{iM} = \left(\frac{M\rho_M}{N}\right)^{1/2} M^{-1/2}\sum_{i=1}^{M}\left(\frac{W_{iM}}{\sqrt{\rho_M}}\right)U_{iM}$$

and, by Lemma A.2, note that $\sqrt{(M\rho_M)/N} \xrightarrow{p} 1$. Therefore, it suffices to show that

$$R_M = M^{-1/2}\sum_{i=1}^{M}\left(\frac{W_{iM}}{\sqrt{\rho_M}}\right)U_{iM} \xrightarrow{d} \mathcal{N}\left(0, [\sigma_U^2 + (1-\rho)\cdot\kappa_U^2]\right).$$

Now

$$\mathbb{E}\left(R_M\right) = M^{-1/2}\sum_{i=1}^{M}\left(\frac{\mathbb{E}\left(W_{iM}\right)}{\sqrt{\rho_M}}\right)\mathbb{E}\left(U_{iM}\right) = \sqrt{\rho_M}M^{-1/2}\sum_{i=1}^{M}\mathbb{E}\left(U_{iM}\right) = 0$$

and

$$\mathbb{V}\left(R_M\right) = M^{-1}\sum_{i=1}^{M}\mathbb{V}\left[\left(\frac{W_{iM}}{\sqrt{\rho_M}}\right)U_{iM}\right].$$

The variance of each term can be computed as

$$\begin{aligned}\mathbb{V}\left[\left(\frac{W_{iM}}{\sqrt{\rho_M}}\right)U_{iM}\right] &= \mathbb{E}\left[\left(\frac{W_{iM}}{\rho_M}\right)U_{iM}^2\right] - \left\{\mathbb{E}\left[\left(\frac{W_{iM}}{\sqrt{\rho_M}}\right)U_{iM}\right]\right\}^2 \\ &= \mathbb{E}\left(U_{iM}^2\right) - \rho_M[\mathbb{E}(U_{iM})]^2 \\ &= \mathbb{V}(U_{iM}) + (1-\rho_M)[\mathbb{E}(U_{iM})]^2.\end{aligned}$$

Therefore,

$$\mathbb{V}\left(R_M\right) = M^{-1}\sum_{i=1}^{M}\mathbb{V}(U_{iM}) + (1-\rho_M)M^{-1}\sum_{i=1}^{M}[\mathbb{E}(U_{iM})]^2 \to \sigma_U^2 + (1-\rho)\kappa_U^2.$$

The final step is to show that the double array

$$Q_{iM} = \frac{M^{-1/2}\left[\left(\frac{W_{iM}}{\sqrt{\rho_M}}\right)U_{iM} - \sqrt{\rho_M}\alpha_{iM}\right]}{\sqrt{\sigma_{U,M}^2 + (1-\rho_M)\kappa_{U,M}^2}} = \frac{1}{\sqrt{M\rho_M}}\frac{(W_{iM}U_{iM} - \rho_M\alpha_{iM})}{\sqrt{\sigma_{U,M}^2 + (1-\rho_M)\kappa_{U,M}^2}},$$

where $\alpha_{iM} = \mathbb{E}(U_{iM})$, satisfies the Lindeberg condition, as in Davidson (1994, Theorem 23.6). Sufficient is the Liapunov condition

$$\sum_{i=1}^{M}\mathbb{E}(|Q_{iM}|^{2+\delta}) \to 0 \text{ as } M \to \infty.$$

Now the term $\sqrt{\sigma_{U,M}^2 + (1-\rho_M)\kappa_{U,M}^2}$ is bounded below by a strictly positive constant because $\sigma_{U,M}^2 \to \sigma_U^2 > 0$. Further,by the triangle inequality,

$$\begin{aligned}\left\{\mathbb{E}\left[\left|W_{iM}U_{iM} - \rho_M\alpha_{iM}\right|^{2+\delta}\right]\right\}^{1/(2+\delta)} &\leq \left[\mathbb{E}\left(W_{iM}\right)\mathbb{E}(\left|U_{iM}\right|^{2+\delta})\right]^{1/(2+\delta)} + \rho_M\left|\alpha_{iM}\right| \\ &\leq \left[\rho_M^{1/(2+\delta)} + \rho_M\right]D_1\end{aligned}$$

where D_1 is constant. Because $\rho_M \in [0,1]$, $\rho_M^{1/(2+\delta)} \geq \rho_M$, and so

$$\mathbb{E}\left[\left|W_{iM}U_{iM} - \rho_M\alpha_{iM}\right|^{2+\delta}\right] \leq \rho_M D_2.$$

Therefore, the Liapunov condition is met if

$$\sum_{i=1}^{M}\frac{\rho_M}{\left(\sqrt{M\rho_M}\right)^{2+\delta}} = \frac{M\rho_M}{(M\rho_M)^{1+(\delta/2)}} = (M\rho_M)^{-\delta/2} \to 0,$$

which is true because $\delta > 0$ and $M\rho_M \to \infty$. We have shown that

$$M^{-1/2}\sum_{i=1}^{M}\frac{\left[\left(\frac{W_{iM}}{\sqrt{\rho_M}}\right)U_{iM} - \sqrt{\rho_M}\alpha_{iM}\right]}{\sqrt{\sigma_{U,M}^2 + (1-\rho_M)\kappa_{U,M}^2}} \xrightarrow{d} \mathcal{N}(0,1)$$

and so, with $\sqrt{\sigma_{U,M}^2 + (1-\rho_M)\kappa_{U,M}^2} \to \sqrt{\sigma_U^2 + (1-\rho)\kappa_U^2}$,

$$M^{-1/2}\sum_{i=1}^{M}\left[\left(\frac{W_{iM}}{\sqrt{\rho_M}}\right)U_{iM} - \sqrt{\rho_M}\alpha_{iM}\right] \xrightarrow{d} \mathcal{N}\left(0, \left[\sigma_U^2 + (1-\rho)\kappa_U^2\right]\right). \square$$

Proof of Lemma 8: This follows directly from Lemma A.4. $\square$

Proof of Theorem 1: We prove part (i), as it is the most important. The other two parts follow similar arguments. To show (i), it suffices to prove two claims. First,

$$\frac{1}{N}\sum_{i=1}^{M}W_{iM}\begin{pmatrix} Z_{iM} \\ X_{iM}\end{pmatrix}\begin{pmatrix} Z_{iM} \\ X_{iM}\end{pmatrix}' - \Gamma \xrightarrow{p} 0 \tag{A.2}$$

holds by Lemma A.3 and the comment following it. The second claim is

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{M}W_{iM}\begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0\end{pmatrix}, \Delta_V + (1-\rho)\Delta_E\right). \tag{A.3}$$

If both claims hold then

$$\sqrt{N}\begin{pmatrix} \hat{\theta}_{\text{ols}} - \theta_M^{\text{causal}} \\ \hat{\gamma}_{\text{ols}} - \gamma_M^{\text{causal}} \end{pmatrix} = \left[\frac{1}{N}\sum_{i=1}^{M} W_{iM}\begin{pmatrix} Z_{iM} \\ X_{iM}\end{pmatrix}\begin{pmatrix} Z_{iM} \\ X_{iM}\end{pmatrix}'\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{M} W_{iM}\begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM}\end{pmatrix}$$
$$= \Gamma^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{M} W_{iM}\begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM}\end{pmatrix} + o_p(1)$$

and then we can apply the continuous convergence theorem and Lemma A.4. The first claim follows from Lemma A.3 and the comment following. For the second claim, we use Lemma A.4 along with the Cramér-Wold device. For a nonzero vector λ, define the scalar

$$U_{iM} = \lambda'\begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM}\end{pmatrix},$$

Given Assumptions 8–10, all of the conditions of Lemma A.4 are met for $\{U_{iM} : i = 1, ..., M\}$. Therefore,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{M} W_{iM}U_{iM} \xrightarrow{d} \mathcal{N}\left(0, [\sigma_U^2 + (1-\rho)\kappa_U^2]\right)$$

where

$$\sigma_U^2 = \lim_{M\to\infty} M^{-1}\sum_{i=1}^{M}\mathbb{V}(U_{iM}) = \lambda'\left\{\lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\mathbb{V}\begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM}\end{pmatrix}\right\}\lambda = \lambda'\Delta_V\lambda$$
$$\kappa_U^2 = \lambda'\left\{\lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\left[\mathbb{E}\begin{pmatrix} X_i\varepsilon_i \\ Z_i\varepsilon_i\end{pmatrix}\right]\left[\mathbb{E}\begin{pmatrix} X_i\varepsilon_i \\ Z_i\varepsilon_i\end{pmatrix}\right]'\right\}\lambda = \lambda'\Delta_E\lambda$$

and so

$$[\sigma_U^2 + (1-\rho)\kappa_U^2] = \lambda'[\Delta_V + (1-\rho)\Delta_E]\lambda$$

By assumption this variance is strictly postive for all $\lambda \neq 0$, and so the Cramér-Wold Theorem proves the second claim. The theorem now follows. □

Proof of Theorem 2: For simplicity, let $\tilde{\theta}_M$ denote θ_M^{causal} and similarly for $\tilde{\gamma}_M$. Then $\tilde{\theta}_M$ and $\tilde{\gamma}_M$ solve the set of equations

$$\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})\tilde{\theta}_M + \mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})\tilde{\gamma}_M = \mathbb{E}(\boldsymbol{X}'\boldsymbol{Y})$$
$$\mathbb{E}(\boldsymbol{Z}'\boldsymbol{X})\tilde{\theta}_M + \boldsymbol{Z}'\boldsymbol{Z}\tilde{\gamma}_M = \mathbb{E}(\boldsymbol{Z}'\boldsymbol{Y}),$$

where we drop the M subscript on the matrices for simplicity. Note that $\boldsymbol{Z}$ is nonrandom and that all moments are well defined by Assumption 9. Multiply the second set of equations by $\mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ to get

$$\mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathbb{E}(\boldsymbol{Z}'\boldsymbol{X})\tilde{\theta}_M + \mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})\tilde{\gamma}_M = \mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathbb{E}(\boldsymbol{Z}'\boldsymbol{Y})$$

and subtract from the first set of equations to get

$$[\mathbb{E}(\boldsymbol{X}'\boldsymbol{X}) - \mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathbb{E}(\boldsymbol{Z}'\boldsymbol{X})]\tilde{\theta}_M = \mathbb{E}(\boldsymbol{X}'\boldsymbol{Y}) - \mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathbb{E}(\boldsymbol{Z}'\boldsymbol{Y})$$

Now, under Assumption 11,

$$\boldsymbol{Y} = \boldsymbol{Y}(0) + \boldsymbol{X}\theta$$

and so

$$\begin{aligned}\mathbb{E}(\boldsymbol{X}'\boldsymbol{Y}) &= \mathbb{E}[\boldsymbol{X}'\boldsymbol{Y}(0)] + \mathbb{E}(\boldsymbol{X}'\boldsymbol{X})\theta \\ \mathbb{E}(\boldsymbol{Z}'\boldsymbol{Y}) &= \boldsymbol{Z}'\boldsymbol{Y}(0) + \mathbb{E}(\boldsymbol{Z}'\boldsymbol{X})\theta\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{E}(\boldsymbol{X}'\boldsymbol{Y}) - \mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathbb{E}(\boldsymbol{Z}'\boldsymbol{Y}) &= \mathbb{E}[\boldsymbol{X}'\boldsymbol{Y}(0)] + \mathbb{E}(\boldsymbol{X}'\boldsymbol{X})\theta \\ &\quad -\mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}(0) - \mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathbb{E}(\boldsymbol{Z}'\boldsymbol{X})\theta \\ &= [\mathbb{E}(\boldsymbol{X}'\boldsymbol{X}) - \mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathbb{E}(\boldsymbol{Z}'\boldsymbol{X})]\theta + \mathbb{E}\{\boldsymbol{X}'[\boldsymbol{Y}(0) - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}(0)]\} \\ &= [\mathbb{E}(\boldsymbol{X}'\boldsymbol{X}) - \mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathbb{E}(\boldsymbol{Z}'\boldsymbol{X})]\theta + \mathbb{E}\{\boldsymbol{X}'[\boldsymbol{Y}(0) - \boldsymbol{Z}\gamma_M]\}\end{aligned}$$

The second term is $\sum_{i=1}^{M} \mathbb{E}_{\boldsymbol{X}} \{X_{iM} [Y_{iM}(0) - Z_{iM}'\gamma_M]\}$, which is zero by Assumption 12. So we have shown that

$$[\mathbb{E}(\boldsymbol{X}'\boldsymbol{X}) - \mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathbb{E}(\boldsymbol{Z}'\boldsymbol{X})]\tilde{\theta}_M = [\mathbb{E}(\boldsymbol{X}'\boldsymbol{X}) - \mathbb{E}(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathbb{E}(\boldsymbol{Z}'\boldsymbol{X})]\theta$$

and solving gives $\tilde{\theta}_M = \theta$. Invertibility holds for M sufficiently large by Assumption 10. Plugging $\tilde{\theta}_M = \theta$ into the orginal second set of equations gives

$$\mathbb{E}(\boldsymbol{Z}'\boldsymbol{X})\theta + \boldsymbol{Z}'\boldsymbol{Z}\tilde{\gamma}_M = \boldsymbol{Z}'\boldsymbol{Y}(0) + \mathbb{E}(\boldsymbol{Z}'\boldsymbol{X})\theta$$

and so $\tilde{\gamma}_M = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}(0) = \gamma_M$. □

Proof of Theorem 3:
By the Frisch-Waugh Theorem (for example, Hayashi, 2000, page 73) we can write

$$\hat{\theta}_{\text{ols}} = \left[N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)(X_{iM} - Z_{iM}\hat{\Pi}_M)'\right]^{-1} N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)Y_{iM}$$

where $Y_{iM} = Y_{iM}(X_{iM})$ and

$$\hat{\Pi}_M = \left(N^{-1}\sum_{i=1}^{M} W_{iM}Z_{iM}Z_{iM}'\right)\left(N^{-1}\sum_{i=1}^{M} W_{iM}Z_{iM}X_{iM}'\right)$$

Plugging in for $Y_{iM} = Z_{iM}'\gamma_M + X_{iM}'\theta + \varepsilon_{iM}$ gives

$$\begin{aligned}N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)Y_{iM} &= N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)X_{iM}'\theta + N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)\varepsilon_{iM} \\ &= \left[N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)(X_{iM} - Z_{iM}\hat{\Pi}_M)'\right]\theta \\ &\quad + N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)\varepsilon_{iM}\end{aligned}$$

where we use the fact that

$$N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)Z_{iM}' = 0$$

by definition of $\hat{\Pi}_M$. It follows that

$$\sqrt{N}\left(\hat{\theta}_{\text{ols}} - \theta\right) = \left[N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)(X_{iM} - Z_{iM}\hat{\Pi}_M)'\right]^{-1} N^{-1/2}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)\varepsilon_{iM}.$$

Now

$$\begin{aligned}
N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)(X_{iM} - Z_{iM}\hat{\Pi}_M)' &= N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)(X_{iM} - Z_{iM}\Lambda_M)' \\
&= N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\Lambda_M)(X_{iM} - Z_{iM}\Lambda_M)' \\
&\quad + N^{-1}\sum_{i=1}^{M} W_{iM}Z_{iM}(\hat{\Pi}_M - \Lambda_M)(X_{iM} - Z_{iM}\Lambda_M)' \\
&= N^{-1}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\Lambda_M)(X_{iM} - Z_{iM}\Lambda_M)' + o_p(1)
\end{aligned}$$

because $\hat{\Pi}_M - \Lambda_M = o_p(1)$ and $N^{-1}\sum_{i=1}^{M} W_{iM}Z_{iM}(X_{iM} - Z_{iM}\Lambda_M)' = O_p(1)$. Further,

$$N^{-1/2}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\hat{\Pi}_M)\varepsilon_{iM} = N^{-1/2}\sum_{i=1}^{M} W_{iM}(X_{iM} - Z_{iM}\Lambda_M)\varepsilon_{iM} + o_p(1)$$

because $N^{-1/2}\sum_{i=1}^{M} W_{iM}Z_{iM}\varepsilon_{iM} = O_p(1)$ by the convergence to multivariate normality.
Next, if we let

$$\dot{X}_{iM} = X_{iM} - Z_{iM}\Lambda_M$$

then we have shown

$$\sqrt{N}\left(\hat{\theta}_{\text{ols}} - \theta\right) = \left(N^{-1}\sum_{i=1}^{M} W_{iM}\dot{X}_{iM}\dot{X}_{iM}'\right)^{-1} N^{-1/2}\sum_{i=1}^{M} W_{iM}\dot{X}_{iM}\varepsilon_{iM} + o_p(1)$$

Now we can apply Theorems 1 and 2 directly. Importantly, ε_{iM} is nonstochastic and so

$$\mathbb{E}(\dot{X}_{iM}\varepsilon_{iM}) = \mathbb{E}(\dot{X}_{iM})\varepsilon_{iM} = 0$$

because

$$\mathbb{E}(\dot{X}_{iM}) = \mathbb{E}(X_{iM}) - Z_{iM}\Lambda_M = 0$$

by Assumption 13. We have already assumed that W_{iM} is independent of $\dot{X}_{iM}$. Therefore, using Theorem 2, we conclude that

$$\sqrt{N}\left(\hat{\theta}_{\text{ols}} - \theta\right) \xrightarrow{d} \mathcal{N}(0, \Gamma_{\dot{X}}^{-1}\Delta_{\text{ehw},\dot{X}}\Gamma_{\dot{X}}^{-1})$$

where

$$\begin{aligned}\Gamma_{\dot{X}} &= \lim_{M\to\infty} M^{-1}\sum_{i=1}^{N}\mathbb{E}\left(\dot{X}_{iM}\dot{X}_{iM}'\right)\\ \Delta_{\text{ehw},\dot{X}} &= \lim_{M\to\infty} M^{-1}\sum_{i=1}^{N}\mathbb{E}\left(\varepsilon_{iM}^2\dot{X}_{iM}\dot{X}_{iM}'\right).\end{aligned}$$

□

Appendix B: 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 2.3. Using a simple parametric model we show that in a Bayesian approach the same issues arise in the choice of estimand. Viewing it from this 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 paper; notation and assumptions introduced in this subsection apply only within this subsection.

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

$$f(\boldsymbol{Y}(\text{low})_M, \boldsymbol{Y}(\text{high})_M, \boldsymbol{X}_M) = \prod_{i=1}^{M} f(Y_i(\text{low}), Y_i(\text{high}), X_i|\beta).$$

Inference then proceeds by specifying a prior distribution for β, say $p(\beta)$.

Let us make this specific, and use the following model. The X_i and W_i are assumed to have binomial distributions with parameters q and ρ,

$$\text{pr}(X_i = \text{high}|Y_i(\text{low}), Y_i(\text{high}), W_i) = q, \qquad \text{pr}(W_i = 1|Y_i(\text{low}), Y_i(\text{high})) = \rho.$$

The pairs $(Y_i(\text{low}), Y_i(\text{high}))$ are assumed to be jointly normally distributed:

$$\begin{pmatrix} Y_i(\text{low}) \\ Y_i(\text{high}) \end{pmatrix}\Bigg| \mu(\text{low}), \mu(\text{high}), \sigma^2(\text{low}), \sigma^2(\text{high}), \kappa \sim \mathcal{N}\left(\begin{pmatrix} \mu(\text{low}) \\ \mu(\text{high}) \end{pmatrix}, \begin{pmatrix} \sigma^2(\text{low}) & \kappa\sigma(\text{low})\sigma(\text{high}) \\ \kappa\sigma(\text{low})\sigma(\text{high}) & \sigma^2(\text{high}) \end{pmatrix}\right),$$

so that the full parameter vector is $\beta = (q, \rho, \mu(\text{low}), \mu(\text{high}), \sigma^2(\text{low}), \sigma^2(\text{high}), \kappa)$.

We change the observational scheme slightly from the previous section to allow for the analytic derivation of posterior distributions. For all units in the population we observe the pair (W_i, X_i), and for units with $W_i = 1$ we observe the outcome $Y_i = Y_i(X_i)$. Define $\tilde{Y}_i = W_i\cdot Y_i$, so we can think of observing for all units in the population the triple $(W_i, X_i, \tilde{Y}_i)$. Let $\boldsymbol{W}_M$, $\boldsymbol{X}_M$, and $\tilde{\boldsymbol{Y}}_M$ be the M vectors of these variables. As before, $\overline{Y}^{\text{obs}}_{\text{high}}$ denotes the average of Y_i in the subpopulation with $W_i = 1$ and $X_i = 1$, and $\overline{Y}^{\text{obs}}_{\text{low}}$ denotes the average of Y_i in the subpopulation with $W_i = 1$ and $X_i = 0$.

The issues studied in this paper arise in this Bayesian approach in the choice of estimand. The descriptive estimand is

$$\theta_M^{\text{descr}} = \frac{1}{M_{\text{high}}} \sum_{i=1}^{M} X_i \cdot Y_i - \frac{1}{M_{\text{low}}} \sum_{i=1}^{M} (1 - X_i) \cdot Y_i.$$

The causal estimand is

$$\theta_M^{\text{causal}} = \frac{1}{M} \sum_{i=1}^{M} \Big(Y_i(\text{high}) - Y_i(\text{low}) \Big).$$

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

$$\theta_\infty^{\text{causal}} = \mu(\text{high}) - \mu(\text{low}).$$

In principle these three estimands are distinct, with their own posterior distributions, but in some cases, notably when M is large, the three posterior distributions are similar.
For each of the three estimands we evaluate the posterior distribution in a special case. In many cases there will not be an analytic solution. However, it is instructive to consider a very simple case where analytic solutions are available. Suppose $\sigma^2(\text{low})$, $\sigma^2(\text{high})$, κ and q are known, so that the only unknown parameters are the two means $\mu(\text{low})$ and $\mu(\text{high})$. Finally, let us use independent, diffuse (improper), prior distributions for μ_{low} and $\mu(\text{high})$.
Then, a standard result is that the posterior distribution for $(\mu_{\text{low}}, \mu(\text{high}))$ given $(\boldsymbol{W}_M, \boldsymbol{X}_M, \tilde{\boldsymbol{Y}}_M)$ is

$$\begin{pmatrix} \mu(\text{low}) \\ \mu(\text{high}) \end{pmatrix} \Big| \, \boldsymbol{W}_M, \boldsymbol{X}_M, \tilde{\boldsymbol{Y}}_M \sim \mathcal{N}\left(\begin{pmatrix} \overline{Y}_{\text{low}}^{\text{obs}} \\ \overline{Y}_{\text{high}}^{\text{obs}} \end{pmatrix}, \begin{pmatrix} \sigma^2(\text{low})/N_{\text{low}} & 0 \\ 0 & \sigma^2(\text{high})/N_{\text{high}} \end{pmatrix} \right).$$

This directly leads to the posterior distribution for $\theta_\infty^{\text{causal}} = \mu(\text{high}) - \mu(\text{low})$:

$$\theta_\infty^{\text{causal}} | \boldsymbol{W}_M, \boldsymbol{X}_M, \tilde{\boldsymbol{Y}}_M \sim \mathcal{N}\left(\overline{Y}_{\text{high}}^{\text{obs}} - \overline{Y}_{\text{low}}^{\text{obs}}, \frac{\sigma^2(\text{low})}{N_{\text{low}}} + \frac{\sigma^2(\text{high})}{N_{\text{high}}} \right).$$

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

$$\theta_M^{\text{descr}} | \boldsymbol{W}_M, \boldsymbol{X}_M, \tilde{\boldsymbol{Y}}_M \sim$$
$$\mathcal{N}\left(\overline{Y}_{\text{high}}^{\text{obs}} - \overline{Y}_{\text{low}}^{\text{obs}}, \frac{\sigma^2(\text{low})}{N_{\text{low}}} \cdot \left(1 - \frac{N_{\text{low}}}{M_{\text{low}}} \right) + \frac{\sigma^2(\text{high})}{N_{\text{high}}} \cdot \left(1 - \frac{N_{\text{high}}}{M_{\text{high}}} \right) \right).$$

The implied posterior interval for θ_M^{descr} is very similar to the corresponding confidence interval based on the normal approximation to the sampling distribution for $\overline{Y}_{\text{high}}^{\text{obs}} - \overline{Y}_{\text{low}}^{\text{obs}}$. If $M_{\text{low}}, M_{\text{high}}$ are large, this posterior distribution converges to

$$\theta_M^{\text{descr}} | \boldsymbol{W}_M, \boldsymbol{X}_M, \tilde{\boldsymbol{Y}}_M, M_{\text{low}} \to \infty, M_{\text{high}} \to \infty \sim \theta_\infty^{\text{causal}} | \boldsymbol{W}_M, \boldsymbol{X}_M, \tilde{\boldsymbol{Y}}_M.$$

If, on the other hand, $N_{\text{low}} = M_{\text{low}}$ and $N_{\text{high}} = M_{\text{high}}$, then the distribution becomes degenerate:

$$\theta_M^{\text{descr}} | \boldsymbol{W}_M, \boldsymbol{X}_M, \tilde{\boldsymbol{Y}}_M, N_{\text{low}} = M_{\text{low}}, N_{\text{high}} = M_{\text{high}} \sim \mathcal{N}\left(\overline{Y}_{\text{high}}^{\text{obs}} - \overline{Y}_{\text{low}}^{\text{obs}}, 0 \right).$$

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

$$\theta_M^{\text{causal}}|\boldsymbol{W}_M,\boldsymbol{X}_M,\tilde{\boldsymbol{Y}}_M \sim \mathcal{N}\left(\overline{Y}_{\text{high}}^{\text{obs}}-\overline{Y}_{\text{low}}^{\text{obs}},\frac{N_{\text{low}}}{M^2}\sigma^2(\text{high})\cdot(1-\kappa^2)+\frac{N_{\text{high}}}{M^2}\sigma^2(\text{low})\cdot(1-\kappa^2)\right.$$
$$+\frac{M-N}{M^2}\sigma^2(\text{high})+\frac{M-N}{M^2}\sigma^2(\text{low})-2\frac{M-N}{M^2}\kappa\sigma(\text{high})\sigma(\text{low})$$
$$\left.+\frac{\sigma^2(\text{high})}{N_{\text{high}}}\cdot\left(1-\left(1-\kappa\frac{\sigma(\text{low})}{\sigma(\text{high})}\right)\frac{N_{\text{high}}}{M}\right)^2+\frac{\sigma^2(\text{low})}{N_{\text{low}}}\cdot\left(1-\left(1-\kappa\frac{\sigma(\text{high})}{\sigma(\text{low})}\right)\frac{N_{\text{low}}}{M}\right)^2\right).$$

Consider the special case where $\kappa=1$, $\sigma(\text{low})=\sigma(\text{high})$. Then

$$\theta_M^{\text{causal}}|\boldsymbol{W}_M,\boldsymbol{X}_M,\tilde{\boldsymbol{Y}}_M,\kappa=1,\sigma(\text{low})=\sigma(\text{high}) \sim \theta_\infty^{\text{causal}}|\boldsymbol{W}_M,\boldsymbol{X}_M,\tilde{\boldsymbol{Y}}_M.$$

The same limiting posterior distribution applies if M goes to infinity.

$$\theta_M^{\text{causal}}|\boldsymbol{W}_M,\boldsymbol{X}_M,\tilde{\boldsymbol{Y}}_M,M_{\text{low}}\to\infty,M_{\text{high}}\to\infty \sim \theta_\infty^{\text{causal}}|\boldsymbol{W}_M,\boldsymbol{X}_M,\tilde{\boldsymbol{Y}}_M.$$

The point is that if the population is large, relative to the sample, the three posterior distributions 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 $\theta_\infty^{\text{causal}}=\mu_{\text{high}}-\mu_{\text{low}}$ is arguably not appropriate, and the posterior inferences for such estimands will differ from those for other estimands such as θ_M^{causal} or θ_M^{descr}.