Finite Population Causal Standard Errors*

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

April 2014

Preliminary–Comments Welcome

Abstract

When a researcher estimates the parameters of a regression function with state level data, using information on all fifty states, what is the interpretation of the standard errors? Researchers typically report standard errors whose formal justification is based on viewing the data as a random sample from a large population of interest, even in cases where we observe outcomes for most or all units in a population. In this paper we explore alternative interpretations for the uncertainty associated with regression estimates. We focus in particular on the case where at least some parameters of the regression function are intended to capture causal effects, and we show that correct standard errors for causal effects, taking account of the finiteness of the population, can be derived using a generalization of randomization inference. Intuitively, these standard errors capture the fact that we do not observe the counterfactual outcomes for each unit corresponding to alternative levels of treatment. We show that our randomization-based standard errors in general are smaller than the conventional robust standard errors, and in some particular cases agree with them. Our results are a leading example of a more general problem: correct inference requires the researcher to be precise about the reference population that defines the estimand, as well as to specify how the observed sample relates to that population. Causal inference is a leading example of where the answer differs from the conventional one, but there are others.

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

†Professor of Economics, Harvard Kennedy School, 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. The goal is then to estimate the parameters of this relation and use the estimates to conduct inference about the values of the parameters in the population. Parameter 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," using, for example, least squares regression.

In the simplest setting with an observed outcome and no covariates in a cross-section of data, the parameter of interest might simply be the population mean, which is 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 estimation and inference in such contexts relies on the assumption that the observed units are a random sample from a large population, where the parameters in this population are the objects of interest. Uncertainty regarding the parameters of interest arises from the difference between the sample and the population. A 95% confidence interval can be interpreted as telling us that if we sample randomly and repeatedly from the population and construct new estimates for each sample, the estimand should be contained in the confidence interval 95% of the time. In many cases this random sampling perspective is reasonable. 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 many cases of interest where there is no population such that the sample can be viewed *both* as (i) small relative to that population, and (ii) as randomly drawn from it. For example, suppose that the units are all fifty states of the United States, or all the countries in the world. If we observe a cross-section of outcomes in a single year and ask how the average outcome varies with state attributes, 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, they typically still report standard errors using formulas that are formally justified by the assumption that the sample is drawn randomly from an infinite population. A theme in this paper is that this random-sampling-from-a-large-population assumption is often not the right one for the problem at hand.

The general perspective we take is that statistics is all about drawing inferences in environments with missing data. If the researcher sees all relevant data, then there is no need to analyze inference, since any question can be answered by simply doing calculations on the data. Outside of this polar case, the relevant question for analyzing inference is, What data are missing? More precisely, 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.

The conventional approach to estimation and inference for a regression in a cross-section can be interpreted as assuming, implicitly, that (i) the units in the sample were selected at random, (ii) the population of units is large (infinite) relative to the sample, (iii) the variables relevant to the regression are observed, and (iv) there is just one state of the world in the population of interest (the one that was realized in the cross-section). Repeated sampling would draw different units from this large population. These assumptions are not appropriate for many settings and questions. In this paper, we allow for the possibility that the number of units in the population is finite, and that the sample may consist of a small or large fraction of all units in the population. Most importantly, we allow that there may be more than one state of the world of interest, even though in a given sample we observe data from a single state of the world.

We focus on the 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 what the average causal effect of the variable is: e.g., 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 whole population of interest, since by definition we will only observe each unit once at the same

time, either in the state where it is treated or the state where it is not.

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 the distinction between *descriptive* estimands and *causal* estimands is not important for estimation if the treatment (*e.g.,* the set of institutions) is randomly assigned. The distinction is also immaterial for inference if population size is large relative to the sample size. However, and this is a key conceptual contribution of the current paper, 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*.

This focus on causal estimands is often explicit 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.

In the current paper we analyze the implications of focusing on causal estimands in finite populations 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.

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 give some thoughts on this

type of question in Section 6.

We make four formal contributions. 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 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 literature going back to Fisher (1935) and Neyman (1923). The early literature focused on settings with randomized assignment without additional covariates. See Rosenbaum (1995) for a textbook discussion. 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 dependendence 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.

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). They will be given without proof. 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 even 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 $\mathbf{Y}$ denoting the $M-$vector with i^{th} element Y_i. The target, or estimand, is the population mean of Y_i,

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

Notation is important in this paper, and we will use the notation to be precise about which quantities are random and which are not, which, for some of the simple examples, may make part of the notation appear redundant. We index the population quantity here by the population size M because we will consider sequences of experiments with populations of increasing size. In that case we will typically make assumptions that ensure that the sequence $\{\mu_M : M = 1, 2 \ldots\}$ converges to a finite constant μ, but it need not be the case that the population mean is identical for each population in the sequence. Technically, assuming that the sequence converges is not necessary for the proofs of the asymptotic properties, but it simplifies the notation and is practically harmless. The dual notation, μ_M and $\overline{Y}_M^{\text{pop}}$, for the same object, captures the different aspects of the 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 fifty 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 (in which case $W_i = 1$) or not (in which case $W_i = 0$), with $\mathbf{W}$ the M-vector with i^{th} element equal to W_i. We let $\{\rho_M\}$ be a sequence of sampling probabilities, one for each population size M, where $\rho_M \in (0, 1)$. We make the following assumption about the sampling process.

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

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

for all $\mathbf{w}$ *with* $w_i \in \{0, 1\}$*, and all* M*.*

This sampling scheme makes the total sample size, $N = \sum_{i=1}^M W_i$, random. An alternative is to draw a random sample of fixed size. Here we use the random sample size in order to allow for the generalizations we consider later. Often the sample is much smaller than the population, $N << M$, but it may be that the sample is the population ($N = M$).

The natural estimator for μ_M is the simple 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. This estimator is, conditional on $N = \sum_{i=1}^M W_i > 0$, unbiased for the population average μ_M:

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

The subscript $\mathbf{W}$ for the variance and expectations operators captures the fact that these variances and expectations are over the distribution generated by the randomness in the sampling indicators W_i: the M values Y_i are fixed quantities. We are interested in the variance of the estimator $\hat{\mu}_M$ conditional on $N > 0$:

$$\mathbb{V}_{\mathbf{W}}\left(\hat{\mu}_M \middle| N, N > 0\right) = \mathbb{E}_{\mathbf{W}}\left[\left(\hat{\mu}_M - \mu_M\right)^2 \middle| N, N > 0\right] = \mathbb{E}_{\mathbf{W}}\left[\left(\overline{Y}^{\text{sample}} - \overline{Y}_M^{\text{pop}}\right)^2 \middle| N, N > 0\right].$$

Because it is conditional 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:

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

which again is itself a random variable.

Also define the population variance of the Y_i,

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

Here we state, without proof, a well-known result from the survey sampling literature.

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

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

Now let us look at the properties of the normalized variance as a function of ρ_M. Because $\mathbb{V}_{\mathbf{w}}\left(\hat{\mu}_M \middle| N=0\right) = 0$, it follows that

$$\mathbb{V}^{\text{norm}}(\hat{\mu}_M) = \mathbf{1}_{N>0}\cdot\sigma_M^2\cdot\left(1-\frac{N}{M}\right).$$

and therefore

$$\mathbb{V}^{\text{norm}}(\hat{\mu}_M) \leq \sigma_M^2\cdot\left(1-\frac{N}{M}\right).$$

The expectation and variance of N/M are ρ_M and $\rho_M(1-\rho_M)/M$, so that by Chebyshev's inequality

$$\text{pr}\left(\mathbb{V}^{\text{norm}}(\hat{\mu}_M) > \varepsilon\right) \leq \frac{\rho_M\cdot(1-\rho_M)}{\left(1-\rho_M-\varepsilon/(2\sigma_M^2)\right)^2},$$

which we can make arbitrarily small by choosing ρ_M sufficiently close to one. Thus, for a fixed population size, ρ_M approaches one, the actual variance as well as the normalized variance approach zero.

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 and assignment rates.

Assumption 3. (Sampling Rate) *The sampling rate ρ_M satisfies*

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

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

$$\mathbb{V}_{\mathbf{w}}(\hat{\mu}_M | N, N > 0) - \frac{\sigma^2}{N} = O_p(M^{-1}),$$

and (ii), as $M \to \infty$,

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

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 fifty 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

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

The estimand is the difference in the two population means,

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

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

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

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

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

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 $\mathbf{W}$. We condition on N_{coast} and N_{land} being positive.

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

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

Again, as in Lemma 1, as the sample size approaches the population size, for a fixed population size, the variance converges to zero.

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, and in our data set, 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, 1935; 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 fifty states in our example), irrespective of whether that state has a minimum wage higher than the federal one or not. Let $\mathbf{Y}(\text{low})_M$, $\mathbf{Y}(\text{high})_M$, $\mathbf{Y}_M$, and $\mathbf{X}_M$ be the M-vectors with ith element equal to $Y_i(\text{high})$, $Y_i(\text{low})$, Y_i, and X_i respectively.

We now define two distinct estimands or population quantities. The first estimand 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) 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 drives the results in this section and is at the core of the paper. We will argue, first, that in many applications, interest is in causal rather than descriptive estimand. However, many textbook discussions implicitly define the estimands in terms of population moments of observed variables, which would correspond to the descriptive estimand. Second, we will argue that in settings where the sample size is large relative to the population size, the distinction between the causal and descriptive estimands matters. In such settings the researcher therefore needs to be explicit about whether an estimand is causal or descriptive.

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.

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* $q_M \in (0, 1)$,

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

for all $\mathbf{x}$ *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 extending the existing results to settings with additional covariates, and we do so in the next section. For expositional purposes we proceed in this section with this assumption.

To formalize the relation between θ_M^{descr} and θ^{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

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

We also introduce notation for the population variances,

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

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

and

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

Note that $\overline{Y}^{\text{pop}}_M(x)$ and $\sigma^2_M(x)$ are averages and variances over the entire population, in contrast to $\overline{Y}^{\text{pop}}_{x,M}$ and $\sigma^2_{x,M}$ which are the averages and variances over the subpopulation of units with $X_i = x$.

The following lemma describes the relation between the two population quantities. Note that θ^{causal}_M is a fixed quantity given the population, whereas θ^{descr}_M is a random variable because it depends on $\mathbf{X}_M$, which is random by Assumption 4. To stress where the randomness in θ^{descr}_M stems from, we use the subscript $\mathbf{X}$ 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 and* $M_{\text{low}}, M_{\text{high}} > 0$. *Then* (i) *the descriptive estimand is unbiased for the causal estimand,*

$$\mathbb{E}_{\mathbf{X}}[\theta^{\text{descr}}_M | M_{\text{high}}] = \theta^{\text{causal}}_M,$$

and (ii),

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

The variance of $\theta^{\text{causal}}_M - \theta^{\text{descr}}_M$ is randomization-based, that is, based on the randomized assignment of the covariate X_i. It is not based on random sampling, and in fact, it cannot be based on random sampling because there is no sampling at this stage, both θ^{causal} and θ^{descr} are population quantities.

Now let us generalize these results to the case where we only observe some of the units in the population. As before, we assume this is a random sample, but we strengthen this by assuming the sampling is random conditional on $\mathbf{X}$.

Assumption 5. (Random Sampling Without Replacement) *Conditional on* $\mathbf{X}_M$ *the sampling indicators* W_i *are independent and identically distributed with* $\text{pr}(W_i = 1 | \mathbf{X}_M) = \rho_M$ *for all* $i = 1, \ldots, M$.

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

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

where the averages of observed outcomes by treatment status are

$$\overline{Y}^{\text{obs}}_{\text{high}} = \frac{1}{N_{\text{high}}} \sum_{i:X_i=\text{high}} W_i \cdot Y_i, \qquad \overline{Y}^{\text{obs}}_{\text{low}} = \frac{1}{M_{\text{low}}} \sum_{i:X_i=\text{low}} W_i \cdot Y_i,$$

and the subsample sizes are $N_{\text{high}} = \sum_{i=1}^{M} W_i \cdot \mathbf{1}_{X_i=\text{high}}$, and $N_{\text{low}} = \sum_{i=1}^{M} W_i \cdot \mathbf{1}_{X_i=\text{low}}$.

The following result is closely related to results in the causal literature. Because we have random sampling and random assignment, we use both subscripts $\mathbf{W}$ and $\mathbf{X}$ for expectations and variances whenever appropriate.

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

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

(ii)

$$\mathbb{V}_{\mathbf{W},\mathbf{X}}\left(\hat{\theta} - \theta^{\text{causal}}_{M}\middle| N_{\text{low}}, N_{\text{high}}, N_{\text{low}} > 0, N_{\text{high}} > 0\right) = \frac{\sigma^2_M(\text{low})}{N_{\text{low}}} + \frac{\sigma^2_M(\text{high})}{N_{\text{high}}} - \frac{\sigma^2_M(\text{low},\text{high})}{M},$$

(iii)

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

(iv)

$$\mathbb{V}_{\mathbf{W},\mathbf{X}}\left(\hat{\theta} - \theta^{\text{descr}}_{M}\middle| N_{\text{low}}, N_{\text{high}}, N_{\text{low}} > 0, N_{\text{high}} > 0\right) = \frac{\sigma^2_M(\text{low})}{N_{\text{low}}} \cdot \left(1 - \frac{N_{\text{low}}}{M_{\text{low}}}\right) + \frac{\sigma^2_M(\text{high})}{N_{\text{high}}} \cdot \left(1 - \frac{N_{\text{high}}}{M_{\text{high}}}\right),$$

(v)

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

Part (ii) of Lemma 5 is a restatement of results in Neyman (1923). Part (iv) is the same result as in Lemma 3. Parts (ii) and (iv) of the lemma imply that

$$\mathbb{V}_{\mathbf{W},\mathbf{X}}\left(\hat{\theta}-\theta^{\text{causal}}\Big|\, N_{\text{low}}, N_{\text{high}}, N_{\text{low}}>0, N_{\text{high}}>0\right)-\mathbb{V}_{\mathbf{W},\mathbf{X}}\left(\hat{\theta}-\theta^{\text{descr}}\Big|\, N_{\text{high}}, N_{\text{low}}\right)$$

$$=\frac{\sigma^2_M(\text{low})}{M_{\text{low}}}+\frac{\sigma^2_M(\text{high})}{M_{\text{high}}}-\frac{\sigma^2_M(\text{low},\text{high})}{M},$$

and by Lemma 4 that is equal to $\mathbb{V}_{\mathbf{W},\mathbf{X}}\left(\theta^{\text{descr}}-\theta^{\text{causal}}\Big|\, N_{\text{high}}, N_{\text{low}}\right)$ which implies 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 want to 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 $\lim_{M\to\infty}\sigma^2_M(\text{low})=\sigma^2(\text{low})$ and $\lim_{M\to\infty}\sigma^2_M(\text{high})=\sigma^2(\text{high})$. Also define, again under Assumption 6, $\lim_{M\to\infty}\sigma^2_M(\text{low},\text{high})=\sigma^2(\text{low},\text{high})$.

Define the normalized variances for the causal and descriptive estimands,

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

and

$$\mathbb{V}^{\text{norm}}_{\text{descr}}=N\cdot\mathbb{V}\left(\hat{\theta}-\theta^{\text{descr}}\Big|\, N_{\text{high}}, N_{\text{low}}\right).$$

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 4–7 hold. Then* (i)*:*

$$\lim_{M\to\infty} \mathbb{V}_{\mathbf{W},\mathbf{X}}\left(\hat{\theta}-\theta^{\text{causal}}\Big|\, N_{\text{high}}, N_{\text{low}}\right)$$

$$= \lim_{M\to\infty} \mathbb{V}_{\mathbf{W},\mathbf{X}}\left(\hat{\theta}-\theta^{\text{descr}}\Big|\, N_{\text{high}}, N_{\text{low}}\right) = \frac{\sigma^2(\text{low})}{N_{\text{low}}} + \frac{\sigma^2(\text{high})}{N_{\text{high}}},$$

(ii) *as* $M\to\infty$*,*

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

and (iii) *as* $M\to\infty$*,*

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

Part (i) of Lemma 6 is a key insight. It shows that we do not need to be concerned about the difference between the two estimands θ^{causal}_M and θ^{descr}_M in settings where the population is large relative to the sample. It is only in settings with the ratio of the sample size to the population size non-negible, and in particular if the sample size is equal to the population size, that there are substantial differences between the two variances.

Although neither the result for the large population limit of the variance for the causal estimand or the result for the large population limit of the variance for the descriptive estimand are novel, the first result known from the causal literature and the second known from the sample survey literature, the equality of the limiting variances had not been pointed out previously as far as we know.

2.4 Estimating the Variance for the Causal Estimand

In this section we consider estimators for the variances in the third and last example. The leading estimator for variances of regression estimators is the Eicker-Huber-White (EHW) robust variance estimator, which, in this simple example with a single binary regressor, is equal to

$$\hat{\mathbb{V}}_{\text{ehw}} = \frac{s^2_{\text{low}}}{N_{\text{low}}} + \frac{s^2_{\text{high}}}{N_{\text{high}}}, \tag{2.3}$$

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

$$s^2_x = \frac{1}{N_x - 1}\sum_{i:X_i=x} W_i\cdot\left(Y_i(x) - \overline{Y}^{\text{sample}}_{x,M}\right)^2.$$

It is also useful to consider the normalized version of this variance,

$$\hat{\mathbb{V}}_{\text{ehw}}^{\text{norm}} = N \cdot \hat{\mathbb{V}}_{\text{ehw}}.$$

and its probability limit:

$$\mathbb{V}_{\text{ehw}}^{\text{norm}} = \frac{\sigma^2(\text{low})}{1-q} + \frac{\sigma^2(\text{high})}{q}.$$

Lemma 7. (Properties of the EHW Variance Estimator) *Suppose Assumptions 4–7 hold. Then* (*i*),

$$\mathbb{E}_{\mathbf{W},\mathbf{X}}[\hat{\mathbb{V}}_{\text{ehw}} | N_{\text{high}}, N_{\text{low}}] = \frac{\sigma^2(\text{low})}{N_{\text{low}}} + \frac{\sigma^2(\text{high})}{N_{\text{high}}},$$

and (*ii*)

$$\hat{\mathbb{V}}_{\text{ehw}}^{\text{norm}} \xrightarrow{p} \mathbb{V}_{\text{ehw}}^{\text{norm}}.$$

One implication of the first part of this lemma is that this variance estimator over-estimates the variance for the descriptive estimand θ^{descr}, by $\sigma_M^2(\text{low})/M_{\text{low}} + \sigma_M^2(\text{high})/M_{\text{high}}$, and over-estimates the variance for the causal estimand θ^{causal} by $\sigma_M^2(\text{low}, \text{high})/M$.

There are two cases where the bias in the EHW variance estimator vanishes. First, suppose the population is large relative to the population size, so that ρ is small. In general the difference between the probability limit of the normalized EHW variance estimator and the normalized variance relative to the descriptive estimand is

$$\mathbb{V}_{\text{ehw}}^{\text{norm}} - \mathbb{V}_{\text{descr}}^{\text{norm}} = \rho \cdot \sigma^2(\text{low}) + \rho \cdot \sigma^2(\text{high}),$$

which goes to zero as $\rho \to 0$. Similarly, the difference between the probability limit of the normalized EHW variance estimator and the normalized variance relative to the causal estimand is

$$\mathbb{V}_{\text{ehw}}^{\text{norm}} - \mathbb{V}_{\text{causal}}^{\text{norm}} = \rho \cdot \sigma^2(\text{low}, \text{high}),$$

which again goes to zero as $\rho \to 0$.

The second important case is where the causal effect is constant and thus $\sigma^2(\text{low}, \text{high}) = 0$. In that case the EHW variance estimator is valid for the causal estimand, irrespective of the population size.

Thus, in general the EHW variance estimator is conservative. If we are interested in the descriptive estimand this would be easy to fix. All we need to know is the ratio of the sample to the population size, $\hat{\rho}_M = N/M$ and we can adjust the EHW variance estimator by multiplying it by $1-\hat{\rho}_M$. If we are interested in the causal estimand the issue becomes more complicated. The problem is that we cannot estimate $\sigma^2(\text{low}, \text{high})$ consistently. A lower bound on this quantity is zero, but that is not in general a sharp bound. Rewriting this variance as

$$\sigma^2(\text{low}, \text{high}) = \sigma^2(\text{low}) + \sigma^2(\text{high}) - 2 \cdot \rho_{\text{low,high}} \cdot \sigma(\text{low}) \cdot \sigma(\text{high}),$$

where $\rho_{\text{low,high}}$ is the correlation in the population between $Y_i(\text{low})$ and $Y_i(\text{high})$, and keeping in mind that we can consistently estimate $\sigma(\text{low})$ and $\sigma(\text{high})$ shows that a sharper lower bound is obtained by the assumption that $Y_i(\text{low})$ and $Y_i(\text{high})$ are perfectly correlated, leading to

$$\sigma^2(\text{low}) + \sigma^2(\text{high}) - 2 \cdot \sigma(\text{low}) \cdot \sigma(\text{high}).$$

However, we can potentially do slightly better by exploiting the fact that the population correlation may be bounded away from one if the distributions of $Y_i(\text{low})$ and $Y_i(\text{high})$ differ by more than location and scale. Both that adjustment, and the one bound based on the perfect correlation case tend to differ little from zero. In practice, therefore, in this setting with a single binary regressor researchers have used the EHW variance estimator as a conservative estimator for the variance.

In the remainder of the paper we study these questions in more general settings. We show that in general the EHW variance estimator is conservative. We also show that the result that the normalized EHW variance is consistent for the actual variance if there is no treatment effect heterogeneity does not hold in general, that is in settings with attributes. We also show that in general, we can do better than the EHW variance estimator by exploiting the presence of these attributes.

2.5 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 discuss the problem from a Bayesian perspective. 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 illustrates 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 section notation will necessarily be a little different from the rest of the paper; notation and assumptions introduced in this subsection apply only within this subsection.

We view the M-vectors $\mathbf{Y}(\text{low})_M$, $\mathbf{Y}(\text{high})_M$, and $\mathbf{X}_M$ as random variables, some observed and some unobserved. We assume the rows of the $M \times 3$ matrix $[\mathbf{Y}(\text{low})_M, \mathbf{Y}(\text{high})_M, \mathbf{X}_M]$ 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:

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

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 are assumed to have a binomial distribution with parameter q

$$\text{pr}(X_i = \text{high}) = q.$$

The pairs $(Y_i(\text{low}), Y_i(\text{high}))$ are assumed to be independent of X_i and 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 $\beta = (q, \mu_{\text{low}}, \mu_{\text{high}}, \sigma^2_{\text{low}}, \sigma^2_{\text{high}}, \kappa)$.

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 $\mathbf{W}_M$, $\mathbf{X}_M$, and $\tilde{\mathbf{Y}}_M$ 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

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

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, these posterior distributions are similar.

For each of the three estimands we can evaluate the posterior distribution given a sample and a prior distribution. 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 σ^2_{low}, σ^2_{high}, κ and q are known, so that the only unknown parameters are the two means μ_{low} and μ_{high}. Finally, let us use independent diffuse (improper) priors for for μ_{low} and μ_{high}.

Then, a standard result is that the posterior distribution for $(\mu_{\text{low}}, \mu_{\text{high}})$ given $(\mathbf{W}_M, \mathbf{X}_M, \tilde{\mathbf{Y}}_M)$ is

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

This directly leads to the posterior distribution for $\theta_\infty^{\text{causal}}$:

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

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

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

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 $\overline{Y}^{\text{obs}}_{\text{high}} - \overline{Y}^{\text{obs}}_{\text{low}}$. If $M_{\text{low}}, M_{\text{high}}$ are large, this posterior distribution converges to

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

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

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

A somewhat longer calculation for $\theta_M^{\rm causal}$ leads to

$$\begin{aligned}\theta_M^{\rm causal}|\mathbf{W}_M, \mathbf{X}_M, \tilde{\mathbf{Y}}_M \sim \mathcal{N}\Bigg(&\overline{Y}_{\rm high}^{\rm obs} - \overline{Y}_{\rm low}^{\rm obs}, \frac{N_{\rm low}}{M^2}\sigma^2_{\rm high}\cdot(1-\kappa^2) + \frac{N_{\rm high}}{M^2}\sigma^2_{\rm low}\cdot(1-\kappa^2)\\
&+\frac{M-N}{M^2}\sigma^2_{\rm high} + \frac{M-N}{M^2}\sigma^2_{\rm low} - 2\frac{M-N}{M^2}\kappa\sigma_{\rm high}\sigma_{\rm low}\\
&+\frac{\sigma^2_{\rm high}}{N_{\rm high}}\cdot\left(1-\left(1-\kappa\frac{\sigma_{\rm low}}{\sigma_{\rm high}}\right)\frac{N_{\rm high}}{M}\right)^2 + \frac{\sigma^2_{\rm low}}{N_{\rm low}}\cdot\left(1-\left(1-\kappa\frac{\sigma_{\rm high}}{\sigma_{\rm low}}\right)\frac{N_{\rm low}}{M}\right)^2\Bigg).\end{aligned}$$

Consider the special case where $\kappa = 1$, $\sigma_{\rm low} = \sigma_{\rm high}$. Then

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

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

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

The point is that if the population is large, relative to the sample, the three posterior distributions agree. However, if the population is small, they differ, and the researcher needs to be precise in defining the estimand. In such cases simply focusing on the super-population estimand $\theta_\infty^{\rm causal} = \mu_{\rm high} - \mu_{\rm low}$ is arguably not appropriate, and the posterior inferences for such estimands will differ from those for other estimands such as $\theta_M^{\rm causal}$ or $\theta_M^{\rm descr}$.

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 can be 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 vector-valued and continuous or discrete. 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 $\mathbf{Z}_M$ and potential outcomes $\mathbf{Y}_M(x)$, 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 of random sampling from an infinite population. The only stochastic component is the matrix $(\mathbf{X}_M, \mathbf{W}_M)$. When we use expectations and variances, these are the random variables that the expectations are taken over.

For a given population 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 $\mathbf{X}$,

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

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 $\Omega = \Omega_M^{\text{pop}}, \Omega_M^{*,\text{pop}}, \Omega_M^{\text{sample}}$. 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. As is well known, the unique solution, assuming no perfect collinearity in the sample, is

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

We are interested in the asymptotic properties of the least squares estimator under different scenarios.

3.2 Descriptive and Causal Estimands

We now define the descriptive and causal estimands that generalize θ^{causal} and θ^{descr} 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:

$$\begin{pmatrix} \theta_M^{\text{descr}} \\ \gamma_M^{\text{descr}} \end{pmatrix} = \begin{pmatrix} \Omega_{XX,M}^{\text{pop}} & \Omega_{XZ',M}^{\text{pop}} \\ \Omega_{ZX',M}^{\text{pop}} & \Omega_{ZZ',M}^{\text{pop}} \end{pmatrix}^{-1} \begin{pmatrix} \Omega_{XY,M}^{\text{pop}} \\ \Omega_{ZY,M}^{\text{pop}} \end{pmatrix}.$$

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

$$\begin{pmatrix} \theta_M^{\text{causal}} \\ \gamma_M^{\text{causal}} \end{pmatrix} = \begin{pmatrix} \Omega_{XX,M}^{*,\text{pop}} & \Omega_{XZ',M}^{*,\text{pop}} \\ \Omega_{ZX',M}^{*,\text{pop}} & \Omega_{ZZ',M}^{*,\text{pop}} \end{pmatrix}^{-1} \begin{pmatrix} \Omega_{XY,M}^{*,\text{pop}} \\ \Omega_{ZY,M}^{*,\text{pop}} \end{pmatrix}.$$

As before, the causal parameters are nonstochastic.

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, $q_M = \mathbb{E}[\sum_{i=1}^{M} X_i/M]$,

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

The component corresponding to θ_M^{causal} is equal to $\overline{Y}_M^{\text{pop}}(1) - \overline{Y}_M^{\text{pop}}(0)$, which is identical to the causal estimand considered in Section 2.3.

3.3 Population Residuals

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

$$\varepsilon_{iM} = Y_{iM} - X_{iM}'\theta_M^{\text{causal}} - Z_{iM}'\gamma_M^{\text{causal}},$$

and define the following expectations, taken over the assignment of the causal covariates X_{iM}:

$$\mu_{X\varepsilon,iM} = \mathbb{E}_{\mathbf{X}}\left[X_{iM} \cdot \varepsilon_{iM}\right] \text{ and } \mu_{Z\varepsilon,iM} = \mathbb{E}_{\mathbf{X}}\left[Z_{iM} \cdot \varepsilon_{iM}\right].$$

The definitions of these residuals closely mirrors that in conventional regression analyses, but their properties are conceptually different. For example, the residuals need not be stochastic. As we show below, if the regression model is correctly specified as a function of the causal variables, so that $Y_{iM}(x) = Y_{iM}(0) + x'\theta$, and the potential causes X_{iM} are randomly assigned, and there are no attributes, then $\varepsilon_{iM} = Y_{iM}(0)$, which is non-stochastic. If the regression model is not correctly specified and $Y_{iM}(x)$ is not linear in x, the residual will be a nondegenerate function of X_{iM} and generally will be stochastic. (To be clear, the residuals are generally non-zero even though they are non-stochastic–it is just that if the functional form of the model is correctly specified, the (unobservable) difference between potential outcomes with different potential causes is non-stochastic.)

Under the assumptions we make, in particular the assumption that the X_{iM} are jointly independent (but not necessarily identically distributed), the ε_{iM}, and also 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 $\mu_{X\varepsilon,iM}$ and $\mu_{Z\varepsilon,iM}$ will not be zero for all i, although under our assumptions stated below and the definition of the residuals, the averages of these expectations over the population will 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 will 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 will 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 Lindeberg-Feller 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) while making the proofs transparent.

Assumption 9. (Moments) *For all M the expected value $\mu_{k,l,m,M} = \mathbb{E}_{\mathbf{X}}[Y_{iM}^k \cdot X_{iM}^l \cdot Z_{iM}^m]$ 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 the assumption offers little because it changes nothing of substance. We also make a full rank assumption.

Assumption 10. (Covergence of moments) *The sequences* $\mathbf{Y}_M$, $\mathbf{Z}_M$ *and* $\mathbf{X}_M$ *satisfy*

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

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

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

In the general case, this is the object we should consistently estimate (as the sample size grows).

For stating the next assumption, let $\mathbf{W}_M$ be the M-vector of sampling indicators.

Assumption 11. (Random Sampling) *For each* M, $\mathbf{W}_M$ *is independent of* $\mathbf{X}_M$, *and the* W_{iM} *are independent and identically distributed with* $\text{pr}(W_{iM} = 1|\mathbf{X}_M) = \rho_M > 0$ *for all* $i = 1, \ldots, M$.

The random sampling assumption implies that, for $N > 0$,

$$\mathbb{E}_{\mathbf{W}}\left[\hat{\Omega}_M^{\text{sample}}\middle| N\right] = \hat{\Omega}_M^{\text{pop}},$$

although this result is not directly relevant for the asymptotic arguments. Instead, we will show convergence in probability once we restrict the rate at which the sampling probability, ρ_M, can tend to zero.

In the proofs of the main results, we combine Assumptions 8 and 11 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 $i = 1, ..., M$.

Assumption 12. (Sampling Rate) *The sampling rate ρ_M satisfies*

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

The requirement $M \cdot \rho_M \to \infty$ 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: $\rho_M = \mathbb{E}(N)/M \to 0$. Technically we should write N_M as the sample size but we drop the M subscript for notational convenience.

A simple but important implication of Assumption 12 is that

$$\frac{N}{M\rho_M} \xrightarrow{p} 1 \quad \text{as } M \to \infty.$$

The argument is simple because, under Assumption 11, for any positive integer M, $N \sim Binomial(M, \rho_M)$, which implies $\mathbb{E}[N] = M\rho_M$ and $\mathbb{V}(N) = M\rho_M(1 - \rho_M)$. Therefore,

$$\begin{aligned} \mathbb{E}\left[\frac{N}{M\rho_M}\right] &= 1 \\ \mathbb{V}\left(\frac{N}{M\rho_M}\right) &= \frac{M\rho_M(1-\rho_M)}{M^2\rho_M^2} = \frac{(1-\rho_M)}{M\rho_M} \to 0, \end{aligned}$$

and so convergence in probability follows from convergence in mean square. In the proofs below we actually use

$$\frac{M\rho_M}{N} \xrightarrow{p} 1,$$

which follows from Slutsky's Theorem because the reciprocal function is continuous at all nonzero values.

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:

Theorem 1. *Suppose Assumptions 8 to 12 hold. Then* (i)

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

(ii)

$$\begin{pmatrix} \theta_M^{\text{descr}} - \theta \\ \gamma_M^{\text{descr}} - \gamma \end{pmatrix} \xrightarrow{p} 0,$$

and (iii)

$$\begin{pmatrix} \theta_M^{\text{causal}} - \theta \\ \gamma_M^{\text{causal}} - \gamma \end{pmatrix} \longrightarrow 0.$$

Proof: See Appendix.

This result follows fairly directly from the assumptions about the moments and the sequence of populations, although allowing $\rho_M \to 0$ 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 using the entire population, where we normalize by the square root of the population size. The key component is the stochastic behavior of the normalized sum of the product of the residuals and the covariates,

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

In our approach this normalized sum of independent but non-identically distributed terms has mean zero – something we verify below – even though each of the separate terms has a non-zero mean. To conclude that (3.1) has a limiting normal distribution we must apply a CLT for independent double arrays. Here we use the Liapunov CLT as stated in Davidson (1994, Theorem 23.11). Under weak moment conditions it follows that

$$\frac{1}{\sqrt{M}} \sum_{i=1}^{M} \begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM} \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Delta_V),$$

where the limit

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

is assumed to exist. By the independence assumption,

$$\mathbb{V}_{\mathbf{X}} \left(\frac{1}{\sqrt{M}} \sum_{i=1}^{M} \begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM} \end{pmatrix} \right) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{V}_{\mathbf{X}} \begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM} \end{pmatrix}.$$

Notice that the asymptotic variance Δ_V differs from conventional EHW asymptotics for regression analysis where each term in the summand has a zero mean — for example, if we had identically distributed sequences for each M. The standard result is

$$\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM} \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Delta_{\text{ehw}})$$

where

$$\begin{aligned}\Delta_{\text{ehw}} &= \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{\mathbf{X}}\left[\begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM} \end{pmatrix}\begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM} \end{pmatrix}'\right] \\ &= \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{\mathbf{X}}\left[\begin{pmatrix} \varepsilon_{iM}^2 X_{iM}X_{iM}' & \varepsilon_{iM}^2 Z_{iM}X_{iM}' \\ \varepsilon_{iM}^2 X_{iM}Z_{iM}' & \varepsilon_{iM}^2 Z_{iM}Z_{iM}' \end{pmatrix}\right]\end{aligned}$$

is the usual EHW expression. This asymptotic variance omits the outer product term

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

In deriving asymptotic normality of the OLS estimators we will assume convergence of the above matrics.

Assumption 13. (Convergence of Second Moments) *The limits in equations (3.2) and (3.3) exist, and Δ_V is positive definite.*

Example To gain some understanding into the properties of the residuals, let us consider a simple example. Table 1 presents the data for an example with four units in the population. The probability of assignment to the binary treatment is $\text{pr}(X_i = 1) = \text{pr}(X_i = 1) = 1/2$. In the population units 2 and 4 were assigned to the treatment and the others to the control group. The causal parameters are $\theta_M^{\text{causal}} = \overline{Y}_M^{\text{pop}}(1) - \overline{Y}_M^{\text{pop}}(0) = 1$ and $\gamma_M^{\text{causal}} = \overline{Y}_M^{\text{pop}}(0) = 1$.

Now consider unit 1. The residual for this unit is

$$\begin{aligned}\varepsilon_1 &= Y_1(X_1) - \theta_M^{\text{causal}}\cdot X_1 - \gamma_M^{\text{causal}} = Y_1(0) + X_i\cdot(Y_1(1) - Y_1(0)) - X_1 - 1 \\ &= 1 - X_1 - 1 = -X_1.\end{aligned}$$

The expected value and variance of this residual (over the distribution induced by the randomness in X_1) are

$$\mathbb{E}_{\mathbf{X}}[\varepsilon_1] = -1/2, \qquad \mathbb{V}_{\mathbf{X}}(\varepsilon_1) = 1/4.$$

Table 1: Example with Four Units

Unit	$Y_i(0)$	$Y_i(1)$	X_i	Y_i	ε_i	$\mathbb{E}_X[\varepsilon_i]$	$\mathbb{E}_X[X_i \cdot \varepsilon_i]$	$\mathbb{V}_X(\varepsilon_i)$	$\mathbb{V}_X(X_i \cdot \varepsilon_i)$
1	1	1	0	1	0	-1/2	-1/2	1/4	1/4
2	0	5	1	5	3	1	2	4	4
3	2	0	0	2	1	-1/2	-3/2	9/4	9/4
4	1	2	1	2	0	0	0	0	0

□

Let us return to the general case and show that the average (taken over the finite population) of the expectations (taken over the distribution of $\mathbf{X}_M$) of the product between the regressors and residuals is zero, even though each of the unit-level expectations may differ from zero:

$$
\begin{aligned}
\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{\mathbf{X}}\left[\begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM} \end{pmatrix}\right] &= \mathbb{E}_{\mathbf{X}}\left[\frac{1}{M}\sum_{i=1}^{M}\begin{pmatrix} X_{iM}\varepsilon_{iM} \\ Z_{iM}\varepsilon_{iM} \end{pmatrix}\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\frac{1}{M}\sum_{i=1}^{M}\begin{pmatrix} X_{iM}(Y_{iM} - X_{iM}'\theta_M^{\text{causal}} - Z_{iM}'\gamma_M^{\text{causal}}) \\ Z_{iM}(Y_{iM} - X_{iM}'\theta_M^{\text{causal}} - Z_{iM}'\gamma_M^{\text{causal}}) \end{pmatrix}\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\frac{1}{M}\sum_{i=1}^{M}\begin{pmatrix} X_{iM}Y_{iM} \\ Z_{iM}Y_{iM} \end{pmatrix}\right] - \mathbb{E}_{\mathbf{X}}\left[\frac{1}{M}\sum_{i=1}^{M}\begin{pmatrix} X_{iM}X_{iM}'\theta_M^{\text{causal}} + X_{iM}Z_{iM}'\gamma_M^{\text{causal}} \\ Z_{iM}X_{iM}'\theta_M^{\text{causal}} + Z_{iM}Z_{iM}'\gamma_M^{\text{causal}} \end{pmatrix}\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\begin{pmatrix} \hat{\Omega}_{XY,M}^{\text{pop}} \\ \hat{\Omega}_{ZY,M}^{\text{pop}} \end{pmatrix}\right] - \mathbb{E}_{\mathbf{X}}\left[\begin{pmatrix} \hat{\Omega}_{XX',M}^{\text{pop}} & \hat{\Omega}_{XZ',M}^{\text{pop}} \\ \hat{\Omega}_{ZX',M}^{\text{pop}} & \hat{\Omega}_{ZZ',M}^{\text{pop}} \end{pmatrix}\right]\begin{pmatrix} \theta_M^{\text{causal}} \\ \gamma_M^{\text{causal}} \end{pmatrix} \\
&= \begin{pmatrix} \Omega_{XY,M}^{\text{pop}} \\ \Omega_{ZY,M}^{\text{pop}} \end{pmatrix} - \begin{pmatrix} \Omega_{XX',M}^{\text{pop}} & \Omega_{XZ',M}^{\text{pop}} \\ \Omega_{ZX',M}^{\text{pop}} & \Omega_{ZZ',M}^{\text{pop}} \end{pmatrix}\begin{pmatrix} \theta_M^{\text{causal}} \\ \gamma_M^{\text{causal}} \end{pmatrix} = 0,
\end{aligned}
$$

by the definition of θ_M^{causal} and γ_M^{causal}.

The main result of our paper allows for random sampling from the population, so we standardize the estimators by the square root of the sample size, N.

Theorem 2. *Suppose Assumptions 8 to 13 hold. Then* (i)

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

(ii)

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

and (iii)

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

Proof: See Appendix.

How do the results in this theorem compare to the standard EHW results on robust variances? 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 $\Gamma^{-1}\Delta_{\text{ehw}}\Gamma^{-1}$, which is the standard EHW variance expression. Moreover, the difference between the two estimands, 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 variance is positive semi-definite, with the difference equal to $\rho \cdot \Delta_E$.

3.6 The Variance when the Regression Function is Correctly Specified

In this section we study the case where the regression function, as a function of the potential cause X_{iM}, is correctly specified. By "correct specification" we mean the following.

Assumption 14. (Linearity of Potential Outcomes) *The potential outcomes satisfy*

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

Assumption 14 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 $\{(Z_{iM}, Y_{iM}(0) : i = 1, 2, ..., M\}$. To this end, define the vector of slope coefficients from the population regression $Y_{iM}(0)$ on Z_{iM}, $i = 1, 2, ..., M$, as

$$\gamma_M = \left(\frac{1}{M}\sum_{i=1}^{M} Z_{iM}Z_{iM}'\right)^{-1}\left(\frac{1}{M}\sum_{i=1}^{M} Z_{iM}Y_{iM}(0)\right). \tag{3.4}$$

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

Assumption 15. (Orthgonality of Assignment) *With γ_M defined above in (3.4)*

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

for all M.

This assumption says that the mean of X_{iM} is orthogonal to the population residuals $Y_{iM}(0) - Z_{iM}'\gamma_M$, which measures the part of $Y_{iM}(0)$ not explained by Z_{iM}. An important special case of this assumption is when $\mathbb{E}_{\mathbf{X}}[X_{iM}]$ is a linear function of Z_{iM}, say $\mathbb{E}_{\mathbf{X}}[X_{iM}] = \Lambda_M Z_{iM}$, $i = 1, ..., M$, for some matrix Λ_M. It is easily seen that Assumption 15 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 15 allows X_{iM} to be systematically related to Z_{iM}, and even related to $Y_{iM}(0)$, provided the mean 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} is restricted; the rest of the distributions are unrestricted.

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

Theorem 3. *Suppose Assumptions 8, 9, 14, and 15 hold. Then for all M,*

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

Proof: See Appendix.

Given Assumptions 14 and 15 we can immediately apply the result from Theorem 2 with θ instead of θ_M, and we also have a simple interpretation for γ_M^{causal}.

A key implication of Assumptions 14 and 15 is that the population residual ε_{iM} is no longer stochastic:

$$\begin{aligned} \varepsilon_{iM} &= Y_{iM}(X_{iM}) - X_{iM}'\theta - Z_{iM}'\gamma_M \\ &= Y_i(0) + X_{iM}'\theta - X_{iM}'\theta - Z_{iM}'\gamma_M \\ &= Y_i(0) - Z_{iM}'\gamma_M. \end{aligned}$$

which does not involve the stochastic components $\mathbf{X}_M$ or $\mathbf{W}_M$. This leads to simplifications in the variance components. The Γ component remains unchanged, but under Assumption 14, Δ_V simplifies to

$$\Delta_V = \begin{pmatrix} \lim_{M\to\infty} \mathbb{V}_{\mathbf{X}} \left(\frac{1}{\sqrt{M}} \sum_{i=1}^{M} X_{iM}\varepsilon_{iM} \right) & 0 \\ 0 & 0 \end{pmatrix} \tag{3.5}$$

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

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 16. (Linearity of the Treatment In Attributes) *For some $K \times J$ matrix Λ_M,*

$$\mathbb{E}_{\mathbf{X}}[X_{iM}] = \Lambda_M Z_{iM}, i = 1, ..., M$$

Recall that this assumption implies Assumption 15, and so we know least squares consistently estimates θ, and it has a limiting normal distribution when scaled by $\sqrt{N}$. But with Assumption 16 we can say more. Namely, the usual heteroskedasticity-robust variance matrix formula is asymptotically valid for $\hat{\theta}_{\text{ols}}$ (but remains conservative for $\hat{\gamma}_{\text{ols}}$). Note that by Assumption 10 we know that Λ_M converges as $M \to \infty$. Let Λ denote the limit of Λ_M.

It is informative to sketch the derivation of the asymptotic variance of $\sqrt{N}\left(\hat{\theta}_{\text{ols}} - \theta\right)$ when we add Assumption 16. 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

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

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

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

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 16. 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

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

We have essentially proven the following theorem, which is our second main result.

Theorem 4. *Suppose Assumptions 8 to 16 hold. Then*

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

There are two key conclusions in this theorem. The first is that the asymptotic variance of $\hat{\theta}_{\text{ols}}$ does not depend on the ratio of the sample to the population size, as measured by

$\rho \in [0,1]$. Second, we conclude that the usual EHW variance matrix is correct for $\hat{\theta}_{\text{ols}}$, and it can be obtained, as in standard asymptotic theory for OLS, 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 (Assumption 14); we have also assumed linearity of the potential cause in the attributes (Assumption 16). 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 follows from the results for randomized experiments. There typically the focus has been on the constant treatment assumption, which is extended to the linearity in Assumption 14. Generally, if linearity holds and X_{iM} is randomized then the conclusions of Theorem 4 hold.

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

4 Estimating the Variance

In this section, for notational simplicity we drop the M subscript on the variables and the parameters. We first study the Eicker-Huber-White robust variance matrix estimator in the context of the linear regression model

$$Y_i = X_i'\beta + \varepsilon_i.$$

The EHW variance is

$$\mathbb{V}_{\text{ehw}} = \left(\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\left[X_iX_i'\right]\right)^{-1}\left(\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\left[(Y_i - X_i'\beta)^2X_iX_i'\right]\right)\left(\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\left[X_iX_i'\right]\right)^{-1}$$

and its estimated counterpart is

$$\hat{\mathbb{V}}_{\text{ehw}} = \left(\frac{1}{N}\sum_{i=1}^{M}W_i\cdot X_iX_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{M}W_i\cdot (Y_i - X_i'\hat{\beta})^2X_iX_i'\right)\left(\frac{1}{N}\sum_{i=1}^{M}W_i\cdot X_iX_i'\right)^{-1}.$$

Now let us turn to the problem of estimating the components of the variance for our descriptive and causal estimands. It is straightforward to estimate Γ. We simply use the average of the matrix of outerproducts over the sample:

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

Also Δ_{ehw} is easy to estimate:

$$\hat{\Delta}_{\text{ehw}} = \hat{\mathbb{E}}\left[\begin{pmatrix} X_i\varepsilon_i \\ Z_i\varepsilon_i \end{pmatrix}\begin{pmatrix} X_i\varepsilon_i \\ Z_i\varepsilon_i \end{pmatrix}'\right] = \frac{1}{N}\sum_{i=1}^{M} W_i \cdot \begin{pmatrix} X_i\hat{\varepsilon}_i \\ Z_i\hat{\varepsilon}_i \end{pmatrix}\begin{pmatrix} X_i\hat{\varepsilon}_i \\ Z_i\hat{\varepsilon}_i \end{pmatrix}'. \tag{4.1}$$

To estimate Δ_V is more challenging. 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 we used a conservative estimator for the variance, which led us back to the conventional robust variance estimator. Here we can do the same. Because

$$\Delta_V = \mathbb{E}_X\left[\mathbb{V}_X\begin{pmatrix} X_i\varepsilon_i \\ Z_i\varepsilon_i \end{pmatrix}\right] \leq \mathbb{E}_X\left[\begin{pmatrix} X_i\varepsilon_i \\ Z_i\varepsilon_i \end{pmatrix}\begin{pmatrix} X_i\varepsilon_i \\ Z_i\varepsilon_i \end{pmatrix}'\right] = \Delta_{\text{ehw}},$$

we can use the estimator in (4.1) as a conservative estimator for the variance. However, we can do better. Instead of using the average of the outerproduct, we can estimate the conditional variance given the attributes:

$$\Delta_Z = \mathbb{E}\left[\mathbb{V}\left(\begin{pmatrix} X_i\varepsilon_i \\ Z_i\varepsilon_i \end{pmatrix}\,\middle|\, Z_i\right)\right].$$

with

$$\Delta_V \leq \Delta_Z \leq \Delta_{\text{ehw}}.$$

To do so we use the methods developed in Abadie, Imbens and Zheng (2012). Define $\ell_Z(i)$ to be the index of the unit closest to i in terms of Z:

$$\ell_Z(i) = \arg\min_{j\in\{1,\ldots,N\},j\neq i} \|Z_i - Z_j\|.$$

Then:

$$\hat{\Delta}_Z = \frac{1}{2N}\sum_{i=1}^{N}\begin{pmatrix} \hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_Z(i)}X_{\ell_Z(i)} \\ \hat{\varepsilon}_i Z_i - \hat{\varepsilon}_{\ell_Z(i)}Z_{\ell_Z(i)} \end{pmatrix}\begin{pmatrix} \hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_Z(i)}X_{\ell_Z(i)} \\ \hat{\varepsilon}_i Z_i - \hat{\varepsilon}_{\ell_Z(i)}Z_{\ell_Z(i)} \end{pmatrix}',$$

and this is our proposed conservative estimator for Δ_V.

We can do better if we also make Assumptions 14 and 15. In that case we can match on both $\hat{\varepsilon}_i$ and Z_{iM}.

$$\ell_{Z,\varepsilon}(i) = \arg\min_{j\in\{1,\ldots,N\},j\neq i}\left\|\begin{pmatrix} Z_i - Z_j \\ \hat{\varepsilon}_{iM} - \hat{\varepsilon}_{jM} \end{pmatrix}\right\|.$$

Then:

$$\hat{\Delta}_{Z\varepsilon} = \frac{1}{2N}\sum_{i=1}^{N}\begin{pmatrix} \hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_{Z,\varepsilon}(i)}X_{\ell_{Z,\varepsilon}(i)} \\ \hat{\varepsilon}_i Z_i - \hat{\varepsilon}_{\ell_{Z,\varepsilon}(i)}Z_{\ell_{Z,\varepsilon}(i)} \end{pmatrix}\begin{pmatrix} \hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_{Z,\varepsilon}(i)}X_{\ell_{Z,\varepsilon}(i)} \\ \hat{\varepsilon}_i Z_i - \hat{\varepsilon}_{\ell_{Z,\varepsilon}(i)}Z_{\ell_{Z,\varepsilon}(i)} \end{pmatrix}',$$

5 Application

In this section we apply the results from this paper to a real data set. The unit of analysis is a state, with the population consisting of the 50 states. We collected data on the average of the logarithm of yearly earnings by state, and indicators on whether the state had a state minimum wage exceeding the federal minimum wage (high), and whether the state was on the coast or not (coastal). The high state minimum wage indicator is viewed as a potential cause, and the coastal indicator as an attribute.

We estimate the regression

$$Y_i = \gamma_0 + \theta \cdot \text{high}_i + \gamma_1 \cdot \text{coastal}_i + \varepsilon_i.$$

We calculate the standard errors under the assumption of a finite population, under the assumption of a correct specification and allowing for misspecification, and under the assumption of an infinite superpopulation (the conventional robust Eicker-Huber-White standard errors). Table 2 presents the results. For the attributes the choice of population size makes a substantial difference.

Table 2: State Regression

		Standard Error			
	estimate	causal finite pop	causal finite pop robust	descriptive finite pop	descriptive infinite pop
potential causes					
high state min. wage	0.058	(0.040)	(0.039)	(0.0)	(0.040)
attributes					
intercept	10.088	(0.003)	(0.020)	(0.0)	(0.020)
coastal state	0.081	(0.010)	(0.034)	(0.0)	(0.034)

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 2000. 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 2000 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 answer descriptive questions, such as "What was the difference in outcomes between Northern and Southern countries in the year 2000?"

These are not the only questions a researcher could focus on, however, and correspondingly, the sample of countries and potential causes in a given year is not the only possible population of interest. A natural alternative question to consider might be, "What is the expected difference in outcomes between Northern and Southern countries in a future year, such as the year 2001?"

This suggests that the population of interest might include each country in a variety of different states of the world that might be realized in future years. The 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, to calculate confidence intervals, first the researcher must specify which countries are under consideration in 2001. Are we interested in a random sample of countries, all countries, or a particular subset of countries? For the moment, suppose it is the same countries observed in 2000. Second, the researcher must make some assumptions about how state outcomes will vary from year to year.

More formally, a key unknown parameter for such a question is the correlation between a country's outcome in 2000 and its outcomes in 2001 (or some other future year).

Indeed, this observation highlights the fact that the question about 2001 is still not sufficiently precise. What do we wish to hold fixed from year to year? Suppose for the moment that we hold observable attributes fixed, and consider a descriptive estimand, such as the difference between Northern and Southern countries.

Then, in the absence of additional data, to conduct inference about the best linear predictor in the subpopulation just we have defined, a researcher would be forced to make an assumption about the extent to which outcomes change year to year in a state. One polar assumption might be that outcomes are perfectly correlated year to year, so that any variation would be due to

changes in potential causes. Clearly this is an extreme assumption, but in this case, the inference we have analyzed in this paper would be appropriate, since analysis of what was observed this year would be the same as answering questions about a future year. If all countries are observed this year, we know the relevant population difference this year and next year with certainty, and the standard error is zero.

Another extreme assumption would specify that conditional on observable state attributes and potentialy causes, outcomes are independently and identically distributed both across states and over time. Thus, the serial correlation of outcomes within a state would be zero, and we assume that conditional on observeables, the variance of a state's future outcome is equal to the cross-state variance in the year 2000.

In such an example, the sampling can be thought of as stratified sampling from a population consisting of units with fixed attributes and different outcomes due to different states of the world. The sampling is stratified so that we see exactly one observation from each unit. For such a case, since the population is large relative to the sample, the standard EHW standard errors are appropriate.

However, we emphasize that the assumptions required to justify the application of EHW in this setting are strong, and we do not see researchers formally stating a potential population for inference, let alone stating and justifying the underlying assumptions about zero serial correlation. Since those assumptions are probably rarely satisfied in practice, one can imagine that a more satisfying approach brings in prior knowledge for serial correlation parameters or perhaps constructs bounds on confidence intervals. Generally, if future predictions are truly the primary question of interest, it seems prudent to study them in settings with panel data. We leave this direction for future work.

7 Conclusion

In this paper we study the interpretation of standard errors in regression analysis when the sample is equal to or close to the population. The conventional interpretation of the standard errors as reflecting the uncertainty coming from random sampling from a large population does not 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 the conventional standard errors that allows for uncertainty coming from both random sampling and

from random assignment. The standard errors for attributes (as opposed to potential causes) of the units do change 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 hypotheses about those regression models. 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 of the randomization inference approach 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. (): 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.

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.

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

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.

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.

Wooldridge, J.M., (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Appendix

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

Lemma A.1. *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 11 and 12 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 12, $(M\rho_M)/N \xrightarrow{p} 1$. 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 12 $M\rho_M \to \infty$. 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.2. *Suppose Assumptions 8 to 12 hold. Then:*

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

Proof: This follows from the previousl 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.2 implies that

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

Proof of Theorem 1: The first claim follows in a straightforward manner from the assumptions and Lemma A.2 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.3. *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 11 and 12 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, as in Lemma A.1, 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

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

and

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

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,

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

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

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

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

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

Proof of Theorem 2: 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 from Lemma A.1 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

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

and then we can apply the continuous convergence theorem and Lemma A.3. The first claim follows from Lemma A.1 and the comment following. For the second claim, we use Lemma A.3 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 to 13, all of the conditions of Lemma A.3) 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

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

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 3: 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

$$\begin{aligned}
\mathbb{E}(\mathbf{X}'\mathbf{X})\tilde{\theta}_M + \mathbb{E}(\mathbf{X}'\mathbf{Z})\tilde{\gamma}_M &= \mathbb{E}(\mathbf{X}'\mathbf{Y}) \\
\mathbb{E}(\mathbf{Z}'\mathbf{X})\tilde{\theta}_M + \mathbf{Z}'\mathbf{Z}\tilde{\gamma}_M &= \mathbb{E}(\mathbf{Z}'\mathbf{Y}),
\end{aligned}$$

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

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

and subtract from the first set of equations to get

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

Now, under Assumption 14,

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

and so

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

It follows that

$$\mathbb{E}(\mathbf{X}'\mathbf{Y}) - \mathbb{E}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbb{E}(\mathbf{Z}'\mathbf{Y}) = \mathbb{E}[\mathbf{X}'\mathbf{Y}(0)] + \mathbb{E}(\mathbf{X}'\mathbf{X})\theta$$

$$-\mathbb{E}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}(0) - \mathbb{E}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbb{E}(\mathbf{Z}'\mathbf{X})\theta$$

$$= [\mathbb{E}(\mathbf{X}'\mathbf{X}) - \mathbb{E}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbb{E}(\mathbf{Z}'\mathbf{X})]\theta + \mathbb{E}\{\mathbf{X}'[\mathbf{Y}(0) - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}(0)]\}$$

$$= [\mathbb{E}(\mathbf{X}'\mathbf{X}) - \mathbb{E}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbb{E}(\mathbf{Z}'\mathbf{X})]\theta + \mathbb{E}\{\mathbf{X}'[\mathbf{Y}(0) - \mathbf{Z}\gamma_M]\}$$

The second term is $\sum_{i=1}^{M} \mathbb{E}_{\mathbf{X}} \{X_{iM} [Y_{iM}(0) - Z_{iM}'\gamma_M]\}$, which is zero by Assumption 15. So we have shown that

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

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

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

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

To obtain the posterior distribution for θ_M^{descr} we write the estimand in terms of observed and unobserved random variables as

$$\theta_M^{\text{descr}} = \frac{1}{M_1}\sum_{i=1}^{M} W_i \cdot X_i \cdot Y_i - \frac{1}{M_0}\sum_{i=1}^{M} W_i \cdot (1-X_i) \cdot Y_i$$

$$+\frac{1}{M_1}\sum_{i=1}^{M} (1-W_i) \cdot X_i \cdot Y_i - \frac{1}{M_0}\sum_{i=1}^{M} (1-W_i) \cdot (1-X_i) \cdot Y_i.$$

The third and fourth term are unobserved. They are independent. Conditional on μ_0, μ_1, and the data $(\mathbf{W}, \mathbf{X}, \mathbf{Y})$ the posterior distribution of the third term is

$$\frac{1}{M_1}\sum_{i=1}^{M} (1-W_i) \cdot X_i \cdot Y_i | \mu_0, \mu_1, \mathbf{W}, \mathbf{X}, \mathbf{Y} \sim \mathcal{N}\left(\frac{M_1 - N_1}{M_1} \cdot \mu_1, \frac{(M_1-N_1)^2}{M_1} \cdot \sigma_1^2 \right).$$

Integrating over μ_1 leads to

$$\frac{1}{M_1}\sum_{i=1}^{M} (1-W_i) \cdot X_i \cdot Y_i | \mathbf{W}, \mathbf{X}, \mathbf{Y} \sim \mathcal{N}\left(\frac{M_1 - N_1}{M_1} \cdot \overline{Y}_1, \left(\frac{M_1-N_1}{M_1}\right)^2 \cdot \sigma_1^2 + \frac{(M_1-N_1)^2}{M_1} \cdot \sigma_1^2 \right).$$

Thus, adding the first term,

$$\frac{1}{M_1}\sum_{i=1}^{M} X_i \cdot Y_i | \mathbf{W}, \mathbf{X}, \mathbf{Y} \sim \mathcal{N}\left(\overline{Y}_1, \left(\frac{M_1-N_1}{M_1}\right)^2 \cdot \sigma_1^2 + \frac{(M_1-N_1)^2}{M_1} \cdot \sigma_1^2 \right)$$

$$= \mathcal{N}\left(\overline{Y}_1, \frac{\sigma_1^2}{N_1} \cdot \left(1 - \frac{N_1}{M_1}\right) \right)$$

Doing the same calculations for the fourth term, and combining the results, leads to

$$\theta_M^{\text{descr}} | \mu_0, \mu_1, \mathbf{W}_M, \mathbf{X}_M, \tilde{\mathbf{Y}}_M \sim \tag{A.4}$$

$$\mathcal{N}\left(\frac{N_1}{M_1} \cdot \overline{Y}_1 + \frac{M_1 - N_1}{M_1} \cdot \mu_1 - \left(\frac{N_0}{M} \cdot \overline{Y}_0 + \frac{M_0 - N_0}{M_0} \cdot \mu_0 \right), \right.$$

$$\left. \frac{(M_1-N_1)^2}{M_1} \cdot \sigma_1^2 + \frac{(M_0-N_0)^2}{M_0} \cdot \sigma_0^2 \right).$$

$$\theta_M^{\text{descr}} | \mathbf{W}_M, \mathbf{X}_M, \tilde{\mathbf{Y}}_M \sim \mathcal{N}\left(\overline{Y}_1 - \overline{Y}_0, \frac{\sigma_0^2}{N_0} \cdot \left(1 - \frac{N_0}{M_0}\right) + \frac{\sigma_1^2}{N_1} \cdot \left(1 - \frac{N_1}{M_1}\right) \right). \tag{A.5}$$

To obtain the posterior distribution for θ_M^{causal} we again write the estimand in terms of observed and unobserved random variables as

$$\theta_M^{\text{causal}} = \frac{1}{M}\sum_{i=1}^{M} W_i \cdot X_i \cdot Y_i(1) - \frac{1}{M}\sum_{i=1}^{M} W_i \cdot (1-X_i) \cdot Y_i(0)$$

$$+\frac{1}{M}\sum_{i=1}^{M} W_i \cdot (1-X_i) \cdot Y_i(1) - \frac{1}{M}\sum_{i=1}^{M} W_i \cdot X_i \cdot Y_i(0)$$
$$+\frac{1}{M}\sum_{i=1}^{M} (1-W_i) \cdot Y_i(1) - \frac{1}{M}\sum_{i=1}^{M} (1-W_i) \cdot Y_i(0).$$

The first term is non-stochastic:

$$\frac{N_1}{M}\overline{Y}_1 - \frac{N_0}{M}\overline{Y}_0$$

The first part of second term, conditional on the data and μ_0, μ_1, is

$$\frac{1}{M}\sum_{i=1}^{M} W_i \cdot (1-X_i) \cdot Y_i(1) | \mu_0, \mu_1, \mathbf{W}_M, \mathbf{X}_M, \tilde{\mathbf{Y}}_M$$
$$\sim \mathcal{N}\left(\frac{N_0}{M}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_0}(\overline{Y}_0 - \mu_0)\right), \frac{N_0}{M^2}\sigma_1^2 \cdot (1-\rho^2)\right)$$

The second part of second term, conditional on the data and μ_0, μ_1, is

$$\frac{1}{M}\sum_{i=1}^{M} W_i \cdot X_i \cdot Y_i(0) | \mu_0, \mu_1, \mathbf{W}_M, \mathbf{X}_M, \tilde{\mathbf{Y}}_M$$
$$\sim \mathcal{N}\left(\frac{N_1}{M}\left(\mu_0 + \rho\frac{\sigma_0}{\sigma_1}(\overline{Y}_1 - \mu_1)\right), \frac{N_1}{M^2}\sigma_0^2 \cdot (1-\rho^2)\right)$$

The third term,

$$\frac{1}{M}\sum_{i=1}^{M} (1-W_i) \cdot Y_i(1) - \frac{1}{M}\sum_{i=1}^{M} (1-W_i) \cdot Y_i(0) | \mu_0, \mu_1, \mathbf{W}_M, \mathbf{X}_M, \tilde{\mathbf{Y}}_M$$
$$\sim \mathcal{N}\left(\frac{M-N}{M}(\mu_1 - \mu_0), \frac{M-N}{M^2}\sigma_1^2\frac{M-N}{M^2}\sigma_0^2 - 2\frac{M-N}{M^2}\rho\sigma_1\sigma_0\right)$$

So, the whole expression, conditional on data and μ_0, μ_1, has mean

$$\begin{aligned}
&\mathbb{E}[\theta_M^{\text{causal}} | \mu_0, \mu_1, \mathbf{W}_M, \mathbf{X}_M, \tilde{\mathbf{Y}}_M] \\
&\quad = \frac{N_1}{M}\overline{Y}_1 - \frac{N_0}{M}\overline{Y}_0 \\
&\qquad + \frac{N_0}{M}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_0}(\overline{Y}_0 - \mu_0)\right) \\
&\qquad + \frac{N_1}{M}\left(\mu_0 + \rho\frac{\sigma_0}{\sigma_1}(\overline{Y}_1 - \mu_1)\right) \\
&\qquad \frac{M-N}{M}(\mu_1 - \mu_0) \\
&\quad = \mu_1 \cdot \left(\frac{N_0}{M} - \frac{N_1}{M} \cdot \rho\frac{\sigma_0}{\sigma_1} + \frac{M-N}{M}\right)
\end{aligned}$$

$$
\begin{aligned}
&+\mu_0 \cdot \left(\frac{N_1}{M} - \frac{N_0}{M} \cdot \rho \frac{\sigma_1}{\sigma_0} + \frac{M-N}{M} \right) \\
&+\frac{N_1}{M}\overline{Y}_1 - \frac{N_0}{M}\overline{Y}_0 \frac{N_0}{M} \rho \frac{\sigma_1}{\sigma_0}\overline{Y}_0 + \frac{N_1}{M}\rho\frac{\sigma_0}{\sigma_1}\overline{Y}_1 \\
= &\mu_1 \cdot \left(1 - \left(1 - \rho\frac{\sigma_0}{\sigma_1} \right) \frac{N_1}{M} \right) \\
&+\mu_0 \cdot \left(1 - \left(1 - \rho\frac{\sigma_1}{\sigma_0} \right) \frac{N_0}{M} \right) \\
&+\frac{N_1}{M}\overline{Y}_1 - \frac{N_0}{M}\overline{Y}_0 \frac{N_0}{M} \rho \frac{\sigma_1}{\sigma_0}\overline{Y}_0 + \frac{N_1}{M}\rho\frac{\sigma_0}{\sigma_1}\overline{Y}_1
\end{aligned}
$$

and variance

$$
\begin{aligned}
&\mathbb{V}([\theta_M^{\text{causal}}|\mu_0, \mu_1, \mathbf{W}_M, \mathbf{X}_M, \tilde{\mathbf{Y}}_M) \\
&= \frac{N_0}{M^2}\sigma_1^2 \cdot (1-\rho^2) \\
&\quad +\frac{N_1}{M^2}\sigma_0^2 \cdot (1-\rho^2) \\
&\quad +\frac{M-N}{M^2}\sigma_1^2 \frac{M-N}{M^2}\sigma_0^2 - 2\frac{M-N}{M^2}\rho\sigma_1\sigma_0
\end{aligned}
$$

The posterior mean without conditioning on μ_0 and μ_1 is is

$$
\begin{aligned}
&\mathbb{E}[\theta_M^{\text{causal}}|\mathbf{W}_M, \mathbf{X}_M, \tilde{\mathbf{Y}}_M] \\
&= \frac{N_1}{M}\overline{Y}_1 - \frac{N_0}{M}\overline{Y}_0 \\
&\quad +\frac{N_0}{M}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_0}(\overline{Y}_0 - \mu_0) \right) \\
&\quad +\frac{N_1}{M}\left(\mu_0 + \rho\frac{\sigma_0}{\sigma_1}(\overline{Y}_1 - \mu_1) \right) \\
&\quad \frac{M-N}{M}(\mu_1 - \mu_0) \\
&= \overline{Y}_1 - \overline{Y}_0.
\end{aligned}
$$

The posterior variance without conditioning on μ_0 and μ_1 is

$$
\begin{aligned}
&\mathbb{V}([\theta_M^{\text{causal}}|\mathbf{W}_M, \mathbf{X}_M, \tilde{\mathbf{Y}}_M) \\
&= \frac{N_0}{M^2}\sigma_1^2 \cdot (1-\rho^2) \\
&\quad +\frac{N_1}{M^2}\sigma_0^2 \cdot (1-\rho^2) \\
&\quad +\frac{M-N}{M^2}\sigma_1^2 + \frac{M-N}{M^2}\sigma_0^2 - 2\frac{M-N}{M^2}\rho\sigma_1\sigma_0
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{\sigma_1^2}{N_1}\cdot\left(1-\left(1-\rho\frac{\sigma_0}{\sigma_1}\right)\frac{N_1}{M}\right)^2\\
&+\frac{\sigma_0^2}{N_0}\cdot\left(1-\left(1-\rho\frac{\sigma_1}{\sigma_0}\right)\frac{N_0}{M}\right)^2\\
=&\frac{\sigma_1^2}{N_1}+\frac{\sigma_0^2}{N_0}\\
&+....\\
&+\frac{M-N}{M^2}\left(\sigma_1^2+\sigma_0^2-2\rho\sigma_1\sigma_0\right)
\end{aligned}
$$

Consider the special case where $\rho = 1$, $\sigma_0 = \sigma_1$. Then the posterior variance is

$$
\mathbb{V}([\theta_M^{\text{causal}}|\mathbf{W}_M,\mathbf{X}_M,\tilde{\mathbf{Y}}_M)_{\rho=1,\sigma_1=\sigma_0}=\frac{\sigma_1^2}{N_1}+\frac{\sigma_0^2}{N_0}
$$

$$
\theta_M^{\text{causal}}|\mathbf{W}_M,\mathbf{X}_M,\tilde{\mathbf{Y}}_M\sim\mathcal{N}\left(\overline{Y}_1-\overline{Y}_0,\right). \tag{A.6}
$$

The posterior variance without conditioning on μ_{low} and μ_{high} is

$$
\begin{aligned}
&\mathbb{V}([\theta_M^{\text{causal}}|\mathbf{W}_M,\mathbf{X}_M,\tilde{\mathbf{Y}}_M)\\
&=\frac{N_{\text{low}}}{M^2}\sigma_{\text{high}}^2\cdot(1-\rho^2)\\
&\quad+\frac{N_{\text{high}}}{M^2}\sigma_{\text{low}}^2\cdot(1-\rho^2)\\
&\quad+\frac{M-N}{M^2}\sigma_{\text{high}}^2+\frac{M-N}{M^2}\sigma_{\text{low}}^2-2\frac{M-N}{M^2}\rho\sigma_{\text{high}}\sigma_{\text{low}}\\
&\quad+\frac{\sigma_{\text{high}}^2}{N_{\text{high}}}\cdot\left(1-\left(1-\rho\frac{\sigma_{\text{low}}}{\sigma_{\text{high}}}\right)\frac{N_{\text{high}}}{M}\right)^2\\
&\quad+\frac{\sigma_{\text{low}}^2}{N_{\text{low}}}\cdot\left(1-\left(1-\rho\frac{\sigma_{\text{high}}}{\sigma_{\text{low}}}\right)\frac{N_{\text{low}}}{M}\right)^2\\
&=\frac{\sigma_{\text{high}}^2}{N_{\text{high}}}+\frac{\sigma_{\text{low}}^2}{N_{\text{low}}}\\
&\quad+....\\
&\quad+\frac{M-N}{M^2}\left(\sigma_{\text{high}}^2+\sigma_{\text{low}}^2-2\rho\sigma_{\text{high}}\sigma_{\text{low}}\right)
\end{aligned}
$$