Finite Population Causal Standard Errors

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

Incomplete Draft, January 2014

Abstract

When a researcher estimates a regression function with state level data, why are there standard errors that differ from zero? Clearly the researcher has information on the entire population of states. Nevertheless researchers typically report conventional robust standard errors, formally justified by viewing the sample as a random sample from a large population. In this paper we investigate the justification for positive standard errors in cases where the researcher estimates regression functions with data from the entire population. We take the perspective that the regression function is intended to capture causal effects, and that standard errors can be justified using a generalization of randomization inference. We show that these randomization-based standard errors in some cases agree with the conventional robust standard errors, and in other cases are smaller than the conventional ones.

Keywords: Regression Analyses, Standard Errors, Confidence Intervals, Random Sampling, Random Assignment, Finite Population, Large Samples, Potential Outcomes, Causality

---

*We are grateful for comments by Joshua Angrist, Daron Acemoglu, Jim Poterba, and especially for many 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. Electronic correspondence: athey@stanford.edu.
§Professor of Economics, and Graduate School of Business Trust Faculty Fellow 2013-2014, Graduate School of Business, Stanford University, and NBER. Electronic correspondence: imbens@stanford.edu.
¶University Distinguished Professor, Department of Economics, Michigan State University, wooldri1@msu.edu
1 Introduction

It is common in empirical work in economics to specify a parametric relation between observable variables in a population of interest. For estimation and inference it is then assumed that the researcher observes values for these variables for a sample of units randomly drawn from this population. Parameter estimates are based on matching in some fashion the relation between the variables in the population to that in the sample, following what Goldberger (1968) and Manski (1988) call the analogy principle. Uncertainty regarding the parameters of interest, that is, the difference between the parameter estimates and their population counterparts, arises from the difference between the sample and the population. In many cases the random sampling perspective is reasonable. If one analyzes individual level data from the Current Population Survey (CPS), the Panel Study of Income Dynamics (PSID), or other surveys, it is clear that the sample analyzed is only a small subset of the population of interest. Although it is often controversial whether the assumption of a random sample is appropriate, it may be a reasonable working assumption. However, in other cases the random sampling assumption is less appropriate. There often is no population that the sample is drawn from. For example, in analyses where the units are the states of the United states, or the countries of the world, there is no sense that these units can be viewed as a random sample from a larger population. This creates a problem for interpreting the uncertainty in the parameter estimates. If the parameters of interest are defined in terms of observable variables defined on units in the population, then if the sample is equal to the population there should be no uncertainty and standard errors should be equal to zero. However, researchers do not report zero standard errors. Instead they report standard errors formally justified by the assumption that the sample is a finite sample of units drawn randomly from an infinite population.

The current paper addresses this issue by arguing that researchers do not intend to define the parameters of interest in terms of relations that hold in the population between variables whose values the researcher observes for units in their sample. Instead, in many cases researchers are fundamentally interested in causal effects. For example, researchers are not simply interested in the difference between the average outcome for countries with one set of institutions and the average outcome for states with a different set of institutions. Rather, researchers are interested in causal effects of these institutions on the outcomes. Formally, the estimand may be the difference between the average outcome if all countries had been exposed to one set

[1]
of institutions and the average outcome if all countries had been exposed to a different set of institutions. The distinction between these two estimands, the former descriptive, and the latter causal, is not important for estimation if the treatment (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 substantially for inference if the sample and population are of similar size. As a result the researcher should 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 an estimator for this variance.

In the current paper we analyze the implications of focusing on causal estimands in finite populations in general observational studies. We allow for discrete or continuous treatments and for the presence of additional covariates that are potentially correlated with the treatments. We make two formal contributions. First we show that in general, as in the special case that Neyman considers, the conventional robust standard errors associated with the work by Huber (), Eicker (), and White (), are conservative for the standard errors for the estimators for the causal parameters. Second, we show that in the case with additional covariates that are correlated with the treatments one can improve on the conventional variance estimator and we propose estimators for the standard errors that are generally smaller than the conventional robust ones.

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), 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 con-
ditional probability of assignment to a binary treatment. He estimates the effects of interest by minimizing test statistics based on conditional randomization. In the current paper we allow explicitly for general dependence of the assignment mechanism of potential causes (discrete or continuous) on the fixed attributes (discrete or continuous) of the units, thus making the methods applicable to general regression settings.

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 (Neyman, 1923; Rubin, 1974; Holland, 1986; Imbens and Wooldridge, 2008). 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 only includes an intercept. The target, or estimand, is the population mean of some variable $Y_i$,\

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

Notation is important in this paper, and we will attempt to be precise, which, for some of the simple examples, may make the notation seem superfluous. We index the population quantity here by the population size $M$ because we will consider sequences of populations with increasing size. In that case we will typically make assumptions that ensure that the sequence $\mu_M$, $M = 1, \ldots, \infty$, converges to a finite constant, but it need not be the case that the population mean is identical for each population in the sequence. The dual notation, $\mu_M$ and $\overline{Y}_{pop}^M$, for the same object, captures the different aspects of the quantity: on the one hand it is a population quantity, for which we commonly use Greek symbols. On the other hand, because the population is finite, it is a simple average, and the latter notation shows the connection to sample 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 draw a random sample of (random) size $N$ from this population, where the sample size is less than or equal to the population size, $N \leq M$. Often the sample is much smaller than the population, $N \ll M$, but it may be that the sample is the population ($N = M$). For convenience we assume that each unit in the population is sampled independently, with common probability $\rho_M$. This sampling scheme makes the total sample size $N$ random. More common 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.

**Assumption 1. (Random Sampling)** Let $W_i \in \{0, 1\}$ be a binary indicator for the inclusion of unit $i$ in the sample, with $\mathbf{W}$ the $M$-vector of sampling indicators, so that

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

for all $w$.

The natural estimator for $\theta_M$ is the simple sample average:

$$\hat{\mu} = \bar{Y}_{\text{sample}} = \frac{1}{N} \sum_{i=1}^M W_i \cdot Y_i,$$

which, conditional on $N = \sum_{i=1}^M W_i > 0$, is unbiased for the population average $\mu_M$. To be formal, let us define $\hat{\mu} = 0$ if $N = 0$, so $\hat{\mu}$ is always defined. We are interested in the variance of this estimator as an estimator of the population average $\mu_M$:

$$\mathbb{V}_W(\hat{\mu} | N) = \mathbb{E}_W [ (\hat{\mu} - \mu_M)^2 | N ] = \mathbb{E}_W \left[ \left( \bar{Y}_{\text{sample}} - \bar{Y}_{\text{pop}}^M \right)^2 \right] | N.$$  

The subscript $W$ 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 indicator: the $Y_i$ are fixed quantities.

It is also useful to define the normalized variance, that is, the variance normalized by the sample size $N$:

$$\mathbb{V}^{\text{norm}}(\hat{\mu}) = N \cdot \mathbb{V}_W(\hat{\mu} | N).$$

Note that this is a function of the sample size, and therefore is a random variable. Also define the population variance of the $Y_i$,

$$\sigma^2_M = S_{Y,M}^2 = \frac{1}{M-1} \sum_{i=1}^M (Y_i - \bar{Y}_{\text{pop}}^M)^2.$$

Here we state, without proof, a well-known result from the survey sampling literature.
Lemma 1. (Random Sampling) Suppose Assumption 1 holds. Then (i) (exact variance):

$$\mathbb{V}(\hat{\mu} | N) = \sigma^2 \cdot \frac{M - N}{M \cdot N} = \frac{\sigma^2_M}{N} \cdot \left(1 - \frac{N}{M}\right),$$

and, (ii) (approximate normalized variance if population and sample are similar size): as $\rho_M \to 1$,

$$\mathbb{V}^{\text{norm}}(\hat{\mu}) \xrightarrow{p} 0.$$  

The key result is that, for a fixed population size, as the sample size approaches the population size, the variance as well as the normalized variance approach zero, irrespective of the population size.

For the next result we need to make assumptions about sequences of populations with increasing size. These sequences are not stochastic, but we assume that some population averages converge. Let $\mu_{k,M}$ be the $k^{\text{th}}$ 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 $\mu_1, \mu_2$, 

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

Define $\sigma^2 = \mu_2 - \mu_1^2$.

Lemma 2. (Large Populations) Suppose Assumptions 1 and 2 hold. Then: (i) (approximate variance if population is large)

$$\lim_{M \to \infty} \mathbb{V}(\hat{\mu} | N) = \frac{\sigma^2}{N},$$

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

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

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. 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 \( X_i \in \{ \text{coast, land} \} \) denote the geographical status of state \( i \). Define, for \( x = \text{coast, land} \),
the population size \( M_x = \sum_{i=1}^{M} 1_{X_i=x} \), and the population averages and variances
\[
\mu_{x,M} = \bar{Y}_{\text{pop},x,M} = \frac{1}{M_x} \sum_{i:X_i=x} Y_i, \quad \text{and} \quad \sigma^2_{x,M} = S^2_{Y,x,M} = \frac{1}{M_x - 1} \sum_{i:X_i=x} (Y_i - \bar{Y}_{\text{pop},x,M})^2.
\]
The estimand is the difference in the two population means,
\[
\theta_M = \bar{Y}_{\text{pop},\text{coast},M} - \bar{Y}_{\text{pop},\text{land},M},
\]
and the natural estimator for \( \theta_M \) is the difference in sample averages by state type,
\[
\hat{\theta} = \bar{Y}_{\text{sample},\text{coast}} - \bar{Y}_{\text{sample},\text{land}},
\]
where the averages of observed outcomes and sample sizes by type are
\[
\bar{Y}_{\text{sample},x} = \frac{1}{N_x} \sum_{i:X_i=x} W_i \cdot Y_i, \quad N_x = \sum_{i=1}^{M} W_i \cdot 1_{X_i=x},
\]
for \( x = \text{coast, land} \). The estimator \( \hat{\theta} \) can also be thought of as the least squares estimator for \( \theta \)
based on minimizing
\[
\arg\min_{\gamma,\theta} \sum_{i=1}^{M} W_i \cdot (Y_i - \gamma - \theta \cdot 1_{X_i=\text{coast}})^2.
\]
The extension of part (i) of Lemma 1 to this case is fairly immediate.

**Lemma 3. (Random Sampling and Regression)** Suppose Assumption 1 holds. Then
\[
\nabla_{W} \left( \hat{\theta} \bigg| N_{\text{coast}}, N_{\text{land}} \right) = \frac{\sigma^2_{\text{coast},M}}{N_{\text{coast}}} \cdot \left( 1 - \frac{N_{\text{coast}}}{M_{\text{coast}}} \right) + \frac{\sigma^2_{\text{land},M}}{N_{\text{land}}} \cdot \left( 1 - \frac{N_{\text{land}}}{M_{\text{land}}} \right).
\]
Note that we condition on \( N_{\text{coast}} \) and \( N_{\text{land}} \) in this variance. 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

Now 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 the normalized 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 had a state minimum wage exceeded by the federal one. Formally, using the Rubin causal model or potential outcome framework (Neyman, 1935; Rubin, 1974; Imbens and Rubin, 2014), we postulate the existence of two potential outcomes for each state, \( Y_i(\text{low}) \) and \( Y_i(\text{high}) \), for earnings without and with a state minimum wage, with \( Y_i \) the outcome corresponding to the 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 \( i \)th 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

\[
\theta_M^{\text{causal}} = \frac{1}{M} \sum_{i=1}^{M} \left( Y_i(\text{high}) - Y_i(\text{low}) \right). \tag{2.1}
\]

We distinguish 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, $\theta^{\text{causal}}$ and $\theta^{\text{descr}}$, that drives the results in this section and is at the core of the paper. We will argue, first, 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 where the ratio of the sample size to the population size, the researcher therefore needs to be explicit about whether an estimand is causal or descriptive. Second, we will argue that in many applications, interest is in causal rather than descriptive estimands.

Let us start with the second point, the interpretation of the two estimands. 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. Alternatively, the causal estimand can be interpreted as the difference in average outcomes if all states would adopt the regulation versus the average outcome if none of the states would adopt of the regulation. We would argue that in such settings the causal estimand is of more interest than the descriptive estimand.

Now let us study the difference between the two estimands. We assume random assignment of the covariate, $X_i$:

**Assumption 3. (Random Assignment)** The $X_i$ are independent and identically distributed.

$$X_i \perp \perp (Y_i(\text{low}), Y_i(\text{high})).$$

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. Such assumptions are unrealistic in many cases. Often they are viewed as more plausible within homogenous subpopulations defined by observable characteristics 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.

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

$$\bar{Y}_M = \frac{1}{M} \sum_{i=1}^{M} Y_i(x), \quad \text{and} \quad \bar{Y}_{x,M} = \frac{1}{M \cdot X_i = x} \sum_{i}^{M} Y_i(x),$$
where, as before, \( M_x = \sum_{i=1}^{M} 1_{X_i = x} \) is the population size by treatment group. Note that because \( X_i \) is a random variable, \( M_{\text{high}} \) and \( M_{\text{low}} \) are random variables too. We can write the two estimands as

\[
\theta_{\text{causal}} = \overline{Y}_{M}^{\text{pop}}(\text{high}) - \overline{Y}_{M}^{\text{pop}}(\text{low}), \quad \text{and} \quad \theta_{\text{descr}} = \overline{Y}_{\text{high},M}^{\text{pop}} - \overline{Y}_{\text{low},M}^{\text{pop}}.
\]

We also introduce notation for the population variances,

\[
\sigma_{M}^{2}(x) = \frac{1}{M - 1} \sum_{i=1}^{M} (Y_i(x) - \overline{Y}_{M}(x))^2, \quad \text{for } x = \text{low, high},
\]

and

\[
\sigma_{M}^{2}(\text{low, high}) = \frac{1}{M - 1} \sum_{i=1}^{M} (Y_i(\text{high}) - Y_i(\text{low}) - (\overline{Y}_{M}^{\text{pop}}(\text{high}) - \overline{Y}_{M}^{\text{pop}}(\text{low})))^2.
\]

Note that \( \overline{Y}_{M}^{\text{pop}}(\text{low}) \) and \( \overline{Y}_{M}^{\text{pop}}(\text{high}) \) are averages over the entire population, not directly observable averages over the subpopulation of units with \( X_i = \text{low} \) and \( X_i = \text{high} \) respectively.

The following lemma describes the relation between the two population quantities. Note that \( \theta_{M}^{\text{causal}} \) is a fixed quantity given the population, whereas \( \theta_{M}^{\text{descr}} \) is a random variable because it depends on \( \mathbf{X}_M \), which is random by Assumption 3. To stress where the randomness stems from, we will use the subscript \( \mathbf{X} \) on the expectations and variance operators here.

**Lemma 4. (Causal versus Descriptive Estimands)** Suppose Assumption 3 holds and \( M_{\text{low}}, M_{\text{high}} > 0 \). Then (i) the descriptive estimand is unbiased for the causal estimand,

\[
\mathbb{E}_{\mathbf{X}}[\theta_{M}^{\text{descr}} \mid M_{\text{high}}] = \theta_{M}^{\text{causal}},
\]

and (ii),

\[
\nabla_{\mathbf{X}} (\theta_{M}^{\text{descr}} \mid M_{\text{high}}) = \mathbb{E}_{\mathbf{X}} \left[ (\theta_{M}^{\text{descr}} - \theta_{M}^{\text{causal}})^2 \mid M_{\text{high}} \right] = \frac{S_{M}^{2}(\text{low})}{M_{\text{low}}} + \frac{S_{M}^{2}(\text{high})}{M_{\text{high}}} - \frac{\sigma_{M}^{2}(\text{low, high})}{M} \geq 0.
\]

The variance of \( \theta_{M}^{\text{causal}} - \theta_{M}^{\text{descr}} \) 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 \( \theta_{M}^{\text{causal}} \) and \( \theta_{M}^{\text{descr}} \) are population quantities.

[9]
Suppose we have a random sample of size $N$ from this population, with the ratio of the sample size to the population size equal to $\rho_M = N/M \in (0,1]$. We focus on the properties what is essentially the same estimator as before,

$$\hat{\theta} = \bar{Y}_{\text{high}} - \bar{Y}_{\text{low}},$$

where the averages of observed outcomes by treatment status are

$$\bar{Y}_{\text{high}} = \frac{1}{N_{\text{high}}} \sum_{i, X_i = \text{high}} W_i \cdot Y_i, \quad \bar{Y}_{\text{low}} = \frac{1}{M_{\text{low}}} \sum_{i, X_i = \text{low}} W_i \cdot Y_i,$$

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

Before analyzing the properties of this estimator, it is useful to introduce one additional estimand, the sample average treatment effect,

$$\theta_{M, \text{causal, sample}} = \frac{1}{N} \sum_{i=1}^M W_i \cdot (Y_{i, \text{high}} - Y_{i, \text{low}}).$$

We analyze the estimator $\hat{\theta}$ as an estimator of the causal estimands $\theta_{M, \text{causal}}$, $\theta_{M, \text{causal, sample}}$ and as of the descriptive estimand $\theta_{M, \text{descr}}$.

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

**Lemma 5. (Variances and Covariances for Causal and Descriptive Estimands)**

Suppose that Assumptions 1–3 hold. Then: (i)

$$\mathbb{V}_{W,X} (\hat{\theta} - \theta_{M, \text{causal}} \mid N_{\text{high}}, N_{\text{low}}) = \frac{\sigma_M^2(\text{low})}{N_{\text{low}}} + \frac{\sigma_M^2(\text{high})}{N_{\text{high}}} - \frac{\sigma_M^2(\text{low, high})}{M},$$

(ii)

$$\mathbb{V}_{W,X} (\hat{\theta} - \theta_{M, \text{causal, sample}} \mid N_{\text{high}}, N_{\text{low}}) = \frac{\sigma_M^2(\text{low})}{N_{\text{low}}} + \frac{\sigma_M^2(\text{high})}{N_{\text{high}}} - \frac{\sigma_M^2(\text{low, high})}{N},$$

(iii)

$$\mathbb{V}_{W,X} (\hat{\theta} - \theta_{M, \text{descr}} \mid N_{\text{high}}, N_{\text{low}}) = \frac{\sigma_M^2(\text{low})}{N_{\text{low}}} \cdot \left(1 - \frac{N_{\text{low}}}{M_{\text{low}}} \right) + \frac{\sigma_M^2(\text{high})}{N_{\text{high}}} \cdot \left(1 - \frac{N_{\text{high}}}{M_{\text{high}}} \right),$$

(ii)

$$\mathbb{V}_X (\theta_{M, \text{causal, sample}} - \theta_{M, \text{causal}} \mid N_{\text{high}}, N_{\text{low}}) = \frac{\sigma_M^2(\text{low, high})}{N} \cdot \left(1 - \frac{N}{M} \right),$$

[10]
(iv),
\[ \mathbb{V}_{\mathbf{w}, \mathbf{x}} \left( \hat{\theta} - \theta_{\text{causal}}^M \middle| N_{\text{high}}, N_{\text{low}} \right) = \mathbb{V} \left( \hat{\theta} - \theta_{\text{descr}}^M \middle| N_{\text{high}}, N_{\text{low}} \right) + \mathbb{V} \left( \theta_{\text{M}}^\text{descr} - \theta_{M}^\text{causal} \middle| N_{\text{high}}, N_{\text{low}} \right). \]

Part (i) of Lemma 5 is a restatement of results in Neyman (1923). Part (ii) is the same result as in Lemma 3. The first two parts of the lemma imply that
\[ \mathbb{V}_{\mathbf{w}, \mathbf{x}} \left( \hat{\theta} - \theta_{\text{causal}} \middle| N_{\text{high}}, N_{\text{low}} \right) - \mathbb{V}_{\mathbf{w}, \mathbf{x}} \left( \hat{\theta} - \theta_{\text{descr}} \middle| N_{\text{high}}, N_{\text{low}} \right) \]
\[ = \frac{\sigma^2_{M}(\text{low})}{M_{\text{low}}} + \frac{\sigma^2_{M}(\text{high})}{M_{\text{high}}} - \frac{\sigma^2_{M}(\text{low, high})}{M}, \]
and by Lemma 4 that is equal to \( \mathbb{V}_{\mathbf{w}, \mathbf{x}} \left( \theta_{\text{descr}} - \theta_{\text{causal}} \middle| N_{\text{high}}, N_{\text{low}} \right) \) which implies part (iii).

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
\[ \mu_{k,m}^M = \frac{1}{M} \sum_{i=1}^{M} Y_i^k \cdot Y_i^m,(\text{low}). \]
We assume that all (cross-)moments up to second order converge to finite limits.

**Assumption 4. (Sequence of Populations)** For nonnegative integers \( k, m \) such that \( k + m \leq 2 \), and some constants \( \mu_{k,m} \),
\[ \lim_{M \to \infty} \mu_{k,m}^M = \mu_{k,m}. \]

**Lemma 6. (Variances and Covariances for Causal and Descriptive Estimands)** Suppose that Assumptions 1–3 hold. Then: and, (iv), for \( \rho \in (0, 1) \),
\[ \lim_{M \to \infty} \mathbb{V} \left( \hat{\theta} - \theta_{\text{causal}} \middle| N_{\text{high}}, N_{\text{low}}, M_{\text{low}}/M = \rho \right) = \lim_{M \to \infty} \mathbb{V} \left( \hat{\theta} - \theta_{\text{descr}} \middle| N_{\text{high}}, N_{\text{low}}, M_{\text{low}}/M = \rho \right) = \frac{\sigma^2(\text{low})}{N_{\text{low}}} + \frac{\sigma^2(\text{high})}{N_{\text{high}}}, \]
where \( \sigma^2(\text{high}) \) and \( \sigma^2(\text{low}) \) are the limits of \( \sigma^2_{M}(\text{high}) \) and \( \sigma^2_{M}(\text{low}) \) respectively, which are finite by Assumption 4.

Lemma 6 shows that we do not need to be concerned about the difference between \( \theta_{M}^\text{causal} \) and \( \theta_{M}^\text{descr} \) in settings where the population is large relative to the sample. It is only in small population settings, and in particular if the sample is equal to the population, that there are substantial differences between the two estimands.


2.4 Estimating the Variance for the Causal Estimand

There exists no unbiased estimator for the variance for the causal estimand given in part (i) of Lemma 5 in general. Instead, researchers often use

\[
\hat{\sigma}^2 = \frac{s^2_{\text{low}}}{N_{\text{low}}} + \frac{s^2_{\text{high}}}{N_{\text{high}}},
\]

(2.3)

where

\[s^2_x = \frac{1}{M_x - 1} \sum_{i: X_i = x} \left( Y_i(x) - \bar{Y}_{x,M}^{\text{sample}} \right)^2,
\]

which is unbiased for \(\sigma^2_M(x)\), for \(x = \text{low, high}\).

There are two important aspects of this variance estimator. First, \(\hat{\sigma}^2\), originally proposed by Neyman (1923), ignores the last term, minus \(\sigma^2_M(\text{low, high})/M\), and it therefore, because this term is non-positive, over-estimates the variance leading to generally conservative confidence intervals. Only in the case where the treatment effect is constant, this variance estimator is unbiased. Second, this variance estimator is identical to the conventional robust variance estimator for the descriptive estimand, justified by random sampling from a large population. Thus, if a researcher is interested in the causal estimand, in the current setting the fact that the population size is finite can be ignored for the purposes of estimating the variance.

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

This is the main 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 regulation such as the state minimum wage), which can be discrete or continuous, can also be vector-valued, and we 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 correlated.

3.1 Set Up

We denote the potential causes by \(X_{iM}\), and the attributes by \(Z_{iM}\). The vector of attributes \(Z_{iM}\) generally includes an intercept. We assume there exists a set of potential outcomes \(Y_{iM}(x)\), with
the realized and observed outcome equal to \( Y_{iM} = Y_{iM}(X_{iM}) \). We view the potential outcomes \( Y_{iM}(x) \) and the attributes \( Z_{iM} \) as deterministic, and the potential cause \( X_{iM} \) as stochastic. However, unlike a randomized experiment, the potential cause is in general not independent of the potential outcomes or the covariates.

In this case 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 of size \( M \), characterized by sets of covariates or attributes \( Z_{iM} \) and potential outcomes \( Y(x)_{iM} \), as well as on the sequence of assignment mechanisms. For the asymptotics we let the size of the population \( M \) go to infinity, with the sample rate, \( \rho_M \), a function of the population size, allowing for \( \rho_M = 1 \) (the sample is the population), and allowing for \( \rho_M \to 0 \) (random sampling from a large population). In the latter case our results agree with the standard robust variance results. Conditional on the (random) sample size \( N \), the sample itself is a random sample from the population. The sampling indicators are \( W_{iM}, i = 1, \ldots, M \), with \( W_{iM} \) the \( M \)-vector of sampling indicators.

The only stochastic component is the pair of vectors \( (X_{iM}, W_{iM}) \). When we use expectations and variances, these are the random variables that the expectations are taken over.

We estimate a linear regression model

\[
Y_i = \theta' X_i + \gamma' Z_i + \varepsilon_i,
\]

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

\[
(\hat{\theta}_{\text{obs}}, \hat{\gamma}_{\text{obs}}) = \text{arg min}_{\theta, \gamma} \sum_{i=1}^{N} (Y_i - \theta' X_i - \gamma' Z_i)^2.
\]

For a given population define the population moments

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

Below we assume that these expectations are finite. Next, define the estimands

\[
\left( \begin{array}{c} \theta_M \\ \gamma_M \end{array} \right) = \left( \begin{array}{cc} \Omega_{XX,M} & \Omega_{XZ',M} \\ \Omega_{XZ',M} & \Omega_{ZZ',M} \end{array} \right)^{-1} \left( \begin{array}{c} \Omega_{XY,M} \\ \Omega_{ZY,M} \end{array} \right).
\]

Also define \( \varepsilon_{iM} \) be the residual

\[
\varepsilon_{iM} = Y_{iM} - \theta_M' X_{iM} - \gamma_M' Z_{iM},
\]

[13]
and define
\[ \mu_{\varepsilon,iM} = \mathbb{E}_X [\varepsilon_{iM}], \quad \text{and} \quad \mu_{X\varepsilon,iM} = \mathbb{E}_X [X_{iM} \cdot \varepsilon_{iM}] \].

These expectations need not be zero. In fact, they typically will not be. Suppose that in population \( M \) for unit \( i \), \( Y_{iM}(x) = 1 \) (non-stochastic), and that in this population \( \theta_M = \gamma_M = 0 \), then \( \varepsilon_{iM} = 1 \) (nonstochastic), and thus \( \mu_{\varepsilon,iM} = 1 \) for this unit.

### 3.2 Assumptions

A key feature is that we now allow for more complicated assignment mechanisms. Let \( f_{i,M}(x) \) denote the distribution of \( X_{iM} \). 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. We do restrict the variation in these distributions, by assuming that it depends only on the attributes \( Z_{iM} \), not on the potential outcomes \( Y_{iM}(x) \), nor on the sampling indicators \( W_{iM} \).

**Assumption 5. (Assignment Mechanism)** The assignments \( X_{1M}, \ldots, X_{MM} \) are independent, but nonidentically distributed. If \( Z_{iM} = Z_{jM} \), then \( f_{i,M}(x) = f_{j,M}(x) \) for all \( x \).

Next, we assume the second moments of the triple \((Y_i, Z_i, X_i)\) are finite.

**Assumption 6. (Moments)** The expected value \( \mu_{k,l,m,M} = \mathbb{E}_X [Y_{iM}^k \cdot X_{iM}^l \cdot Z_{iM}^m] \) is bounded by \( C_M \) for all nonnegative integers \( k, l, m \) such that \( k + l + m \leq 4 \).

**Assumption 7. (Asymptotic Sequences)** The sequences \( Y_M, Z_M \) and \( X_M \) satisfy
\[
\mathbb{E}_X \left[ \frac{1}{M} \sum_{i=1}^M \begin{pmatrix} Y_{iM} \\ Z_{iM} \\ X_{iM} \end{pmatrix} \begin{pmatrix} Y_{iM} \\ Z_{iM} \\ X_{iM} \end{pmatrix}' \right] \rightarrow \Omega = \begin{pmatrix} \Omega_{YY} & \Omega_{YZ} & \Omega_{YX'} \\ \Omega_{ZY} & \Omega_{ZZ} & \Omega_{ZX'} \\ \Omega_{XY} & \Omega_{XZ'} & \Omega_{XX'} \end{pmatrix},
\]
with \( \Omega \) full rank. In addition the sequence \( C_M \) in Assumption 6 is bounded by a constant \( C \).

If Assumption 7 holds, we can define the limiting values of the finite population estimands \( \theta_M \) and \( \gamma_M \):
\[
\lim_{M \to \infty} \begin{pmatrix} \theta_M \\ \gamma_M \end{pmatrix} = \begin{pmatrix} \theta \\ \gamma \end{pmatrix} = \begin{pmatrix} \Omega_{XX} & \Omega_{XX'} \\ \Omega_{ZX'} & \Omega_{ZZ'} \end{pmatrix}^{-1} \begin{pmatrix} \Omega_{XY} \\ \Omega_{ZY} \end{pmatrix}.
\]

**Assumption 8. (Random Sampling)** Conditional on \( X_M \) the sampling indicators \( W_{iM} \) are independent and identically distributed with \( \Pr(W_{iM} = 1|X_M) = \rho_M \) for all \( i = 1, \ldots, M \).
Assumption 9. (Sampling Rate) The sampling rate $\rho_M$ satisfies

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

This assumption will guarantee that as $M \to \infty$ the sample size $N$ will go to infinity too, but at the same time allow for the possibility that the sample size is a negligible fraction of the population size.

3.3 The General Case

Theorem 1. Suppose Assumptions 5–9 hold. Then

$$\left( \hat{\theta}_{ols} - \theta_M, \hat{\gamma}_{ols} - \gamma_M \right) \overset{p}{\to} 0$$

Proof: See Appendix.

This result follows fairly directly from the assumptions about the moments and the sequence of populations. Next is the first of the two main results in the paper.

Theorem 2. Suppose Assumptions 5–9 hold. Then

$$\sqrt{N} \left( \hat{\theta}_{ols} - \theta_M, \hat{\gamma}_{ols} - \gamma_M \right) \overset{d}{\to} \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Gamma^{-1} (\Delta_V + (1 - \rho) \cdot \mu_E \cdot \mu_E') \Gamma^{-1} \right),$$

where

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

$$\Delta_V = \lim_{M \to \infty} \mathbb{V}_X \left( \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \begin{pmatrix} X_i \epsilon_i \\ Z_i \epsilon_i \end{pmatrix} \right),$$

(3.1)

and

$$\mu_E = \lim_{M \to \infty} \mathbb{E}_X \left[ \frac{1}{M} \sum_{i=1}^{M} \begin{pmatrix} X_i \epsilon_i \\ Z_i \epsilon_i \end{pmatrix} \right].$$

(3.2)

How does this compare to the standard results on robust variances? This case is captured by $\rho = 0$. Define

$$\Delta_E = \lim_{M \to \infty} \mathbb{E}_X \left[ \frac{1}{M} \sum_{i=1}^{M} \begin{pmatrix} X_i \epsilon_i \\ Z_i \epsilon_i \end{pmatrix} \begin{pmatrix} X_i \epsilon_i \\ Z_i \epsilon_i \end{pmatrix} \right],$$

so that $\Delta_E = \Delta_V + \mu_E \cdot \mu_E'$. The asymptotic variance in the case with $\rho = 0$ reduce to $\Gamma^{-1} \Delta_E \Gamma^{-1}$, which is the standard Eicker-Huber-White variance. If the sample size is nonnegligible as a fraction of the population sizes, $\rho > 0$, the difference between the Eicker-Huber-White variance and the finite population variance is positive definite, with the difference equal to $\rho \cdot \mu_E \cdot \mu_E'$. [15]
3.4 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 \( X_i \), is correctly specified.

**Assumption 10. (Linearity of Potential Outcomes)** The potential outcomes satisfy

\[
Y_i(x) = Y_i(0) + \theta' x.
\]

First we establish the relation between the population estimand \( \theta_M \) and the slope of the potential outcome function.

**Theorem 3.** Suppose Assumptions 5–10 hold. Then for all \( M \),

\[
\theta_M = \theta.
\]

Thus, the result from Theorem 2 directly applies with \( \theta \) instead of \( \theta_M \).

A key implication of the correct specification assumption is that the residual \( \varepsilon_{iM} \) is no longer stochastic:

\[
\varepsilon_{iM} = Y_{iM}(X_{iM}) - \theta' X_{iM} - \gamma'_M Z_{iM}
\]

\[
= Y_i(0) + \theta' X_{iM} - \theta' X_{iM} - \gamma'_M Z_{iM}
\]

\[
= Y_i(0) - \gamma'_M Z_{iM},
\]

which does not involve the stochastic components \( X_M \) or \( W_M \). Hence \( \sum_{i=1}^{M} Z_{iM} \cdot \varepsilon_{iM} \) is non-stochastic. This leads to simplifications in the variance components. The \( \Gamma \) component remains unchanged, but under Assumption 10, \( \Delta_V \) simplifies to

\[
\Delta_V = \begin{pmatrix}
\lim_{M \to \infty} V_X 
\left( \frac{1}{\sqrt{M}} \sum_{i=1}^{M} X_i \varepsilon_i \right) & 0 \\
0 & 0
\end{pmatrix}
\]  

(3.3)

and

\[
\mu_E = \lim_{M \to \infty} \mathbb{E}_X \left[ \frac{1}{M} \sum_{i=1}^{M} \left( \frac{X_i \varepsilon_i}{Z_i \varepsilon_i} \right) \right].
\]

(3.4)

Now define the residuals

\[
\hat{X}_i = X_i - Z_i \Omega^{-1}_{ZZ'} \Omega_{XX'}, \quad \hat{Z}_i = Z_i - X_i \Omega^{-1}_{XX'} \Omega_{ZZ'}
\]
and their associated outer products,
\[ \Omega_{\dot{X}\dot{X}'} = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[\dot{X}_i \dot{X}'_i], \quad \Omega_{\dot{Z}\dot{Z}'} = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[\dot{Z}_i \dot{Z}'_i] \]
and finally define the outer products with the residuals \( \varepsilon_i \),
\[ \Omega_{\dot{X}\varepsilon^2\dot{X}'} = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[\dot{X}_i \varepsilon^2_i \dot{X}'_i], \quad \Omega_{\dot{Z}\varepsilon^2\dot{Z}'} = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[\dot{Z}_i \varepsilon^2_i \dot{Z}'_i], \]
\[ \Omega_{\dot{Z}_i\varepsilon^2\dot{X}'} = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[\dot{Z}_i \varepsilon^2_i \dot{X}'_i], \quad \Omega_{\dot{X}_i\varepsilon^2\dot{Z}'} = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[\dot{X}_i \varepsilon^2_i \dot{Z}'_i]. \]

The next theorem contains the second of the main results of the paper.

**Theorem 4.** Suppose Assumptions 5–10. Then
\[ \sqrt{N} \left( \hat{\theta}_{\text{ols}} - \theta - \hat{\gamma}_{\text{ols}} - \gamma_M \right) \xrightarrow{d} \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_{\theta\theta'} & V_{\theta\gamma'} \\ V_{\gamma\theta'} & V_{\gamma\gamma'} \end{pmatrix} \right), \]
where
\[ V_{\theta\theta'} = \Omega_{\dot{X}\dot{X}'}^{-1}, \Omega_{\dot{X}\varepsilon^2\dot{X}'}^{-1}, \Omega_{\dot{X}\dot{X}'}^{-1}, \]
\[ V_{\gamma\gamma'} = \rho \delta \Omega_{\dot{X}\dot{X}'}^{-1}, \Omega_{\dot{X}\varepsilon^2\dot{X}'}^{-1}, \Omega_{\dot{X}\dot{X}'}^{-1}, \delta' + (1 - \rho) \Omega_{\dot{Z}\dot{Z}'}^{-1}, \Omega_{\dot{Z}\varepsilon^2\dot{Z}'}^{-1}, \Omega_{\dot{Z}\dot{Z}'}^{-1}. \]

Note that \( V_{\theta\theta'} \) does not depend on \( \rho \): the conventional robust (Huber-Eicker-White) variance is valid irrespective of the population size. For the case with \( X_i \) binary and no attributes beyond the intercept this result follows from results for randomized experiments. Theorem 4 extends this to general non-experimental settings. The variance on the least squares estimates of the coefficients on the attributes, \( V_{\gamma\gamma'} \), does depend on the ratio of sample to population size. Here the conventional robust variance over-estimates the uncertainty in the estimates.

Three special cases demonstrate the range of the result. If the sample size is negible relative to the population size (\( \rho = 0 \)), then we recover the familiar result for least squares estimators with the Eicker-Huber-White robust variance:
\[ \sqrt{N} \left( \hat{\gamma} - \gamma_M \right) \xrightarrow{d} \mathcal{N} \left( 0, \Omega_{\dot{Z}\dot{Z}'}^{-1}, \Omega_{\dot{Z}\varepsilon^2\dot{Z}'}^{-1}, \Omega_{\dot{Z}\dot{Z}'}^{-1} \right). \]
At the other extreme, if the sample is the population (\( \rho = 1 \)), then:
\[ \sqrt{N} \left( \hat{\gamma} - \gamma_M \right) \xrightarrow{d} \mathcal{N} \left( 0, \delta \Omega_{\dot{X}\dot{X}'}^{-1}, \Omega_{\dot{X}\varepsilon^2\dot{X}'}^{-1}, \Omega_{\dot{X}\dot{X}'}^{-1}, \delta' \right). \]
Finally, if all the covariates are attributes and there are no potential causes, then
\[
\sqrt{N} \left( \hat{\gamma} - \gamma_M \right) \xrightarrow{d} \mathcal{N} \left( 0, (1 - \rho)\Omega_{ZZ}^{-1}\Omega_{Z\varepsilon Z}\Omega_{ZZ'}^{-1} \right).
\]
In that case if the sample is equal to the population ($\rho = 1$) the variance of $\hat{\gamma}$ is equal to zero.

4 Estimating the Variance

It is straightforward to estimate $\Gamma$. We simply use the average of the matrix of outerproducts over the sample:
\[
\hat{\Gamma} = \frac{1}{N} \sum_{i=1}^{M} W_i \cdot \begin{pmatrix} Z_i & X_i \end{pmatrix} \begin{pmatrix} Z_i & X_i \end{pmatrix}^\prime.
\]

The second term in the expression for $\Delta$ in (3.4) is also easy to estimate as the average of the outer products, similar to the corresponding expression in the Eicker-Huber-White variance estimator:
\[
\hat{E} \left[ \begin{pmatrix} X_i \varepsilon_i \\ Z_i \varepsilon_i \end{pmatrix} \begin{pmatrix} X_i \varepsilon_i \\ Z_i \varepsilon_i \end{pmatrix}^\prime \right] = \frac{1}{N} \sum_{i=1}^{M} W_i \cdot \begin{pmatrix} X_i \hat{\varepsilon}_i \\ Z_i \hat{\varepsilon}_i \end{pmatrix} \begin{pmatrix} X_i \hat{\varepsilon}_i \\ Z_i \hat{\varepsilon}_i \end{pmatrix}^\prime.
\]

(4.1)

To estimate the term multiplying $\rho$ in the (3.4) 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
\[
E \left[ V \left( \begin{pmatrix} X_i \varepsilon_i \\ Z_i \varepsilon_i \end{pmatrix} \left| Z_i \right. \right) \right] \leq E \left[ \begin{pmatrix} X_i \varepsilon_i \\ Z_i \varepsilon_i \end{pmatrix} \begin{pmatrix} X_i \varepsilon_i \\ Z_i \varepsilon_i \end{pmatrix}^\prime \right],
\]
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:
\[
E \left[ V \left( \begin{pmatrix} X_i \varepsilon_i \\ Z_i \varepsilon_i \end{pmatrix} \left| Z_i \right. \right) \right].
\]

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:
\[
\hat{E} \left[ \nabla \left( \frac{X_i \varepsilon_i}{Z_i \varepsilon_i} \right) \right] = \frac{1}{2N} \sum_{i=1}^{N} \left( \hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_2(i)} X_{\ell_2(i)} \right) \left( \hat{\varepsilon}_i Z_i - \hat{\varepsilon}_{\ell_2(i)} Z_{\ell_2(i)} \right)',
\]
and this is our proposed estimator for the term multiplying \( \rho \) in (3.4).

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 Huber-Eicker-White standard errors). Table 1 presents the results. For the attributes the choice of population size makes a substantial difference.

6 A Small Simulation Study

To assess the performance of the variance estimators in practice we carry out a small simulation study.

First we generate data for the population. For a given population size \( M \) we generate a two component vector of attributes, with \( Z_{i1} = 1 \), and \( Z_{i2} \) independent and normally distributed with mean zero and unit variance. To generate the outcome at \( x = 0 \) we first draw an independent and normally distributed random variable \( \tilde{\varepsilon}_i \), with mean zero and unit variance. We then take the residual from the projection on \( Z_M \) to generate the \( \varepsilon_i \). Then
\[
Y_i(0) = \gamma_0 + \gamma_1 \cdot Z_{i2} + \varepsilon_i.
\]
Table 1: State Regression

<table>
<thead>
<tr>
<th>potential causes</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>causal estimate</td>
</tr>
<tr>
<td>high state minimum wage</td>
<td>0.058 (0.040)</td>
</tr>
</tbody>
</table>

with $\gamma_0 = \gamma_1 = 1$.

We keep this population with $Y(0)_M$ and $Z_M$ fixed over the repeated samples. In each sample we draw a random sample of size $N$ from this population and assign the potential cause. The scalar potential cause $X_i$ is generated as

$$X_i = Z_i + \eta_i,$$

where $\eta_i \sim N(0,1)$, leading to the observed outcome

$$Y_i = Y_i(0) + \theta \cdot X_i,$$

with $\theta = 1$. Each time we run the regression

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

We calculate for $\theta$ and $\gamma_1$ the standard deviation over the replications, and the finite and infinite population average standard error (fp and robust respectively), and check whether the true values of $\theta$ and $\gamma_1$ are inside the nominal 95% confidence interval. We also calculate the average of the conventional robust standard errors and also check whether the nominal 95% confidence interval based on those standard errors includes the true value of the parameter.

Table 2 presents the results for 10,000 replications, for $N = M = 100$ and for the case with $N = 100, M = 200$. 

[20]
Table 2: Summary Statistics Simulation Study

<table>
<thead>
<tr>
<th>s.d.</th>
<th>se_{fp}</th>
<th>se_{robust}</th>
<th>95% ci_{fp}</th>
<th>95% ci_{robust}</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 100, M = 200</td>
<td>θ</td>
<td>0.109</td>
<td>0.088</td>
<td>0.088</td>
</tr>
<tr>
<td>N = 100, M = 100</td>
<td>γ₁</td>
<td>0.121</td>
<td>0.118</td>
<td>0.141</td>
</tr>
</tbody>
</table>

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. This is only a first step. The same concerns arise in many other settings, and the implications of the randomization inference approach would need to be worked out for those settings.
REFERENCES


