ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

An International Society for the Advancement of Economic Theory in its Relation to Statistics and Mathematics

http://www.econometricsociety.org/

Econometrica, Vol. 88, No. 1 (January, 2020), 265–296

SAMPLING-BASED VERSUS DESIGN-BASED UNCERTAINTY IN REGRESSION ANALYSIS

ALBERTO ABADIE
Massachusetts Institute of Technology and NBER

SUSAN ATHEY
Graduate School of Business, Stanford University and NBER

GUIDO W. IMBENS
Graduate School of Business and Department of Economics, Stanford University and NBER

JEFFREY M. WOOLDRIDGE
Department of Economics, Michigan State University

Econometrica, Vol. 88, No. 1 (January, 2020), 265–296

SAMPLING-BASED VERSUS DESIGN-BASED UNCERTAINTY IN REGRESSION ANALYSIS

ALBERTO ABADIE
Massachusetts Institute of Technology and NBER

SUSAN ATHEY
Graduate School of Business, Stanford University and NBER

GUIDO W. IMBENS
Graduate School of Business and Department of Economics, Stanford University and NBER

JEFFREY M. WOOLDRIDGE
Department of Economics, Michigan State University

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

KEYWORDS: Finite population, potential outcomes, descriptive and causal estimands.

1. INTRODUCTION

IN THE DOMINANT APPROACH to inference in the social sciences, uncertainty about population parameters is induced by random sampling from the population. Moreover, it is typically assumed that the sample comprises only a small fraction of the population of interest. This perspective is a natural and attractive one in many instances. For example, if one analyzes individual-level data from the U.S. Current Population Survey, the

Alberto Abadie: abadie@mit.edu
Susan Athey: athey@stanford.edu
Guido W. Imbens: imbens@stanford.edu
Jeffrey M. Wooldridge: wooldri1@msu.edu

We are grateful for comments by Daron Acemoglu, Joshua Angrist, Matias Cattaneo, Alex Lenk, Jonas Metzger, Evan Munro, Michael Pollmann, Jim Poterba, Karthik Rajkumar, Tymon Słoczyński, Bas Werker, and seminar participants at Microsoft Research, Michigan, Brown University, MIT, Stanford, Princeton, NYU, Columbia, Tilburg University, the Tinbergen Institute, American University, Montreal, Michigan State, Maryland, Pompeu Fabra, Carlos III, and University College London, three referees, and especially for discussions with Gary Chamberlain. An earlier version of this paper circulated under the title "Finite Population Causal Standard Errors" (Abadie, Athey, Imbens, and Wooldridge (2014)).

 https://doi.org/10.3982/ECTA12675

Panel Study of Income Dynamics, or the one percent public-use sample from the U.S. Census, it is natural to regard the sample as a small random subset of the population of interest. In many other settings, however, this sampling perspective is less appropriate. For example, Manski and Pepper (2018) wrote, "Random sampling assumptions, however, are not natural when considering states or counties as units of observation." In this article, we provide an alternative framework for the interpretation of uncertainty in regression analysis regardless of whether a substantial fraction of the population or even the entire population is included in the sample. While our framework accommodates sampling-based uncertainty, it also takes into account design-based uncertainty, which arises when the parameter of interest is defined in terms of the unobserved outcomes that some units would attain under a certain intervention. Design-based uncertainty is often explicitly accounted for in the analysis of randomized experiments where it is the basis of randomization inference (Neyman, (1923/1990), Rosenbaum, (2002), Imbens and Rubin (2015)), but it is rarely explicitly acknowledged in regression analyses or, more generally, in observational studies (exceptions include Samii and Aronow (2012), Freedman (2008), Lu (2016), Lin (2013)).

To illustrate the differences between sampling-based inference and design-based inference, consider two simple examples. In the example of Table I, there is a finite population consisting of n units with each unit characterized by a pair of variables, Y_i and Z_i. Consider an estimand that is a function of the full set of pairs $\{(Y_i, Z_i)\}_{i=1}^n$. Uncertainty about such an estimand arises when we observe the values (Y_i, Z_i) only for a sample, that is, for a subset of the population. In Table I, inclusion of unit i in a sample is coded by the binary variable $R_i \in \{0, 1\}$. An estimator is a function of the observed data, $\{(R_i, R_iY_i, R_iZ_i)\}_{i=1}^n$. Sampling-based inference uses information about the process that determines the sampling indicators $R_1, \ldots, R_n$ to assess the variability of estimators across different samples. The second and third sets of columns in Table I depict such alternative samples. Table II depicts a different scenario in which we observe, for each unit in the population, the value of one of two potential outcome variables, either $Y_i^*(1)$ or $Y_i^*(0)$, but not both. The binary variable $X_i \in \{0, 1\}$ indicates which potential outcome we observe. Consider an estimand that is a function of the full set of triples $\{(Y_i^*(1), Y_i^*(0), X_i)\}_{i=1}^n$. As before, an estimator is a function of the observed data, the pairs (X_i, Y_i), for $i = 1, \ldots, n$, where $Y_i = Y_i^*(X_i)$ is the realized value. Design-based inference uses information about the process that determines the assignments $X_1, \ldots, X_n$ to assess the variability of estimators across different samples. The second and third sets of columns in Table II depict such alternative samples.

More generally, we can have missing data processes that combine features of these two examples, with some units not included in the sample at all, and with some of the variables

TABLE I

SAMPLING-BASED UNCERTAINTY (✓ IS OBSERVED, ? IS MISSING)

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

TABLE II

DESIGN-BASED UNCERTAINTY (✓ IS OBSERVED, ? IS MISSING)

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

not observed for the sampled units. Articulating both the exact nature of the estimand of interest and the source of uncertainty that makes an estimator stochastic is a crucial first step to valid inference. For this purpose, it will be useful to distinguish between descriptive estimands, where uncertainty stems solely from not observing all units in the population of interest, and causal estimands, where the uncertainty stems, at least partially, from unobservability of some of the potential outcomes.

The main formal contribution of this article is to generalize the results for the approximate variance for multiple linear regression estimators associated with the work by Eicker (1967), Huber (1967), and White (1980a, 1980b, 1982), EHW from hereon, in two ways. First, our framework allows for sampling from a finite population, whereas the EHW results assume random sampling from an infinite population. Second, our framework explicitly takes into account design-based uncertainty. Incorporating these generalizations requires developing a new framework for regression analysis, nesting as special cases the Neyman (1923/1990) analysis of randomized experiments with binary treatments, as well as the generalizations to randomized experiments with additional regressors in Samii and Aronow (2012), Freedman (2008), and Lin (2013). We show that in large samples, the widely used EHW robust standard errors are conservative, and only correct in special cases. Moreover, we show that the presence of attributes—that is, characteristics of the units fixed in our repeated sampling thought experiments—can be exploited to improve on the EHW variance estimator, and we propose variance estimators that do so. Another advantage of the formal separation into sampling-based and design-based uncertainty is that it allows us to clarify the distinction between the assumptions needed for internal and external validity (Shadish, Cook, and Campbell (2002), Manski (2013), Deaton (2010)) in terms of these two sources of uncertainty.

Our results are relevant in empirical settings where researchers have a random sample from a finite population and the ratio of the sample size to the population size is sufficiently large so that the proposed finite-population correction matters. Examples of such settings include large-scale experiments (see Muralidharan and Niehaus (2017)), settings where the cost of data acquisition motivates the use of random samples (see, e.g., Keels, Duncan, DeLuca, Mendenhall, and Rosenbaum (2005)), as well as analyses based on public-use census samples, like the 2010 Integrated Public Use Microdata Series (IPUMS) data (which is a 10 percent sample of the U.S. Census). More importantly in our view, our results are relevant in empirical settings where it is not natural to think of the data as a random sample from a well-defined population. Instead, the researcher may have the entire population, for example, states or counties as in the Manski and Pepper (2018) quote, or the set of all visits to a website, or the researcher may have a convenience

sample. In that case, our design-based approach to uncertainty provides a coherent interpretation for sampling-based standard errors. It also provides methods that exploit the presence of attributes to calculate improved (i.e., less conservative) standard errors.

2. A SIMPLE EXAMPLE

In this section, we set the stage for the problems discussed in the current article by discussing least squares estimation in a simple example with a single binary regressor. We make four points. First, we show how design-based uncertainty affects the variance of regression estimators. Second, we show that the standard Eicker–Huber–White (EHW) variance estimator remains conservative when we take into account design-based uncertainty. Third, we show that there is a simple finite-population correction to the EHW variance estimator for descriptive estimands but not for causal estimands. Fourth, we discuss the relation between the two sources of uncertainty and the notions of internal and external validity. Proofs of the results in this section are in the Supplemental Material (Abadie, Athey, Imbens, and Wooldridge (2020)).

We focus on a setting with a finite population of size n. We sample N units from this population, with $R_i \in \{0, 1\}$ indicating whether a unit was sampled ($R_i = 1$) or not ($R_i = 0$), so that $N = \sum_{i=1}^{n} R_i$. There is a single binary regressor, $X_i \in \{0, 1\}$, and n_x (resp. N_x) is the number of units in the population (resp. the sample) with $X_i = x$. Units could be U.S. states and the binary regressor X_i could be an indicator for a state regulation, say the state having a right-to-carry law (RTC), as in Manski and Pepper (2018) and Donohue, Aneja, and Weber (2019). We view the regressor X_i not as a fixed attribute or characteristic of each unit, but as a cause or policy variable whose value could have been different from the observed value. This generates missing data of the type shown in Table II, where only some of the states of the world are observed, implying that there is design-based uncertainty. Formally, using the Rubin causal model or potential outcome framework (Neyman (1923/1990), Rubin (1974), Holland (1986), Imbens and Rubin (2015)), we postulate the existence of two potential outcomes for each unit, denoted by $Y_i^*(1)$ and $Y_i^*(0)$. For the RTC example, $Y_i^*(1)$ and $Y_i^*(0)$ could be state-level crime rates with and without RTC. The realized outcome is

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

which is the observed state-level crime rate in the RTC example.

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

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

of Y_i and X_i for all units in the population. If an estimand cannot be written as a function of $(\mathbf{Y}, \mathbf{X}, \mathbf{R})$ because it depends on the potential outcomes $\mathbf{Y}^*(1)$ and $\mathbf{Y}^*(0)$, then we label it a *causal* estimand.[1]

We now consider in our binary regressor example three closely related estimands, one descriptive and two causal:

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

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

and

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

In this section, we focus on the properties of the difference-in-sample-means estimator:

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

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

Next, we consider the first two moments of $\widehat{\theta}$ under the following two assumptions:

ASSUMPTION 1—Random Sampling/External Validity:

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

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

ASSUMPTION 2—Random Assignment/Internal Validity:

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

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

We start by studying the first moment of the estimator, conditional on (N_1, N_0), and only for the cases where $N_1 \geq 1$ and $N_0 \geq 1$. We leave this latter conditioning implicit in the notation throughout this section. We also condition implicitly on the fixed potential

[1]This does not define an exhaustive partition the set of all possible estimands. For example, there could be estimands that are functions of $(\mathbf{Y}, \mathbf{X}, \mathbf{R})$, but not of $(\mathbf{Y}, \mathbf{X})$, although it is difficult to think of interesting ones that are.

outcomes $\mathbf{Y}^*(1)$ and $\mathbf{Y}^*(0)$. Taking the expectation only over the random sampling, or taking the expectation only over the random assignment, or over both, we find

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

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

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

Next, we look at the variance of the estimator, maintaining both the random assignment and random sampling assumption. Define the population variances

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i^*(x) - \frac{1}{n}\sum_{j=1}^{n} Y_j^*(x)\right)^2, \quad \text{for } x = 0, 1,$$

and

$$S_\theta^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i^*(1) - Y_i^*(0) - \frac{1}{n}\sum_{j=1}^{n}\big(Y_j^*(1) - Y_j^*(0)\big)\right)^2.$$

We consider the variance of $\widehat{\theta}$, as well as two conditional versions of this variance. We define the "sampling variance" conditional on $\mathbf{X}$, so that only the sampling uncertainty is taken into account. Analogously, we define the "design variance" conditional on $\mathbf{R}$, so that only the design uncertainty is taken into account. To make the different variances interpretable, we look at the expected value of the variances, taking the expectation over both the assignment and the sampling:

$$V^{\text{total}}(N_1, N_0, n_1, n_0) = \text{var}(\widehat{\theta}|N_1, N_0) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n_0 + n_1}, \tag{2.3}$$

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

$$V^{\text{design}}(N_1, N_0, n_1, n_0) = E\big[\text{var}(\widehat{\theta}|\mathbf{R}, N_1, N_0)\big|N_1, N_0\big] = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N_0 + N_1}.$$

COMMENT 1—Neyman Variance: The variance $V^{\text{total}}(N_1, N_0, n_1, n_0)$ is the one derived by Neyman (1923/1990) for randomized experiments.

COMMENT 2—Causal versus Descriptive Estimands: In general, the variances $V^{\text{sampling}}(N_1, N_0, n_1, n_0)$ and $V^{\text{design}}(N_1, N_0, n_1, n_0)$ cannot be ranked: the sampling variance can be very close to zero if the sampling rates $\frac{N_0}{n_0}$ and $\frac{N_1}{n_1}$ are close to 1, but it can also be larger than the design variance if the sampling rates are small and the variance of the treatment effect is substantial.

COMMENT 3—Infinite Population Case: For fixed N_0 and N_1, if $n_0, n_1 \to \infty$, the total variance and the sampling variance are equal:

$$\lim_{n_0, n_1, \to \infty} V^{\text{total}}(N_1, N_0, n_1, n_0) = \lim_{n_0, n_1, \to \infty} V^{\text{sampling}}(N_1, N_0, n_1, n_0) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}.$$

This result will be seen to carry over to more general cases in Section 3.

COMMENT 4—Finite Population Correction: Whether the estimand is θ^{causal} or θ^{descr}, ignoring the fact that the population is finite generally leads to an overstatement of the variance on average because it ignores the fact that we observe a non-negligible share of the population:

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

$$V^{\text{sampling}}(N_1, N_0, \infty, \infty) - V^{\text{sampling}}(N_1, N_0, n_1, n_0) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} \geq 0.$$

If the estimand is $\theta^{\text{causal,sample}}$, however, the population size is irrelevant because units in the population but not in the sample do not contribute to the estimand of interest. Then,

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

COMMENT 5—Internal versus External Validity: Often, researchers are concerned about both the internal and external validity of estimands and estimators (Shadish, Cook, and Campbell (2002), Manski (2013), Deaton (2010)). The distinction between sampling- and design-based uncertainty allows us to clarify these concerns. Internal validity bears on the question of whether $E[\widehat{\theta}|\mathbf{R}, N_1, N_0]$ is equal to $\theta^{\text{causal,sample}}$. This relies on random assignment of the treatment. Whether or not the sampling is random is irrelevant for this question because $\theta^{\text{causal,sample}}$ conditions on which units were sampled. External validity bears on the question of whether $E[\theta^{\text{causal,sample}}|N_1, N_0]$ is equal to θ^{causal}. This relies on the random sampling assumption and does not require that the assignment is random. However, for $\widehat{\theta}$ to be a good estimator of θ^{causal}, which is often the most interesting estimand, we need both internal and external validity, and thus both random assignment and random sampling.

In this single binary regressor example, the EHW variance estimator can be written as

$$\widehat{V}^{\text{ehw}} = \frac{N_1 - 1}{N_1^2}\widehat{S}_1^2 + \frac{N_0 - 1}{N_0^2}\widehat{S}_0^2, \quad \text{where } \widehat{S}_1^2 = \frac{1}{N_1 - 1}\sum_{i=1}^{n} R_i X_i \left(Y_i - \frac{1}{N_1}\sum_{i=1}^{n} R_i X_i Y_i\right)^2,$$

and $\widehat{S}_0^2$ is defined analogously. Adjusting the degrees of freedom, using the modification proposed in MacKinnon and White (1985) specialized to this binary regressor example, we obtain $\widetilde{V}^{\text{ehw}} = \widehat{S}_1^2/N_1 + \widehat{S}_0^2/N_0$, which is identical to the variance estimator proposed by Neyman (1923/1990). The expectation of this modified EHW variance estimator $\widetilde{V}^{\text{ehw}}$ (conditional on N_0 and N_1) is equal to the sampling variance in the infinite population case, $V^{\text{sampling}}(N_1, N_0, \infty, \infty)$.

We could also estimate the variance using resampling methods, which would give us variance estimates close to $\widehat{V}^{\text{ehw}}$. To be precise, suppose we use the bootstrap where we draw N_1 bootstrap observations from the N_1 treated units and N_0 bootstrap units from the N_0 control units. In that case, the bootstrap variance would in expectation (over the bootstrap replications) be equal to $\widehat{V}^{\text{ehw}}$.

COMMENT 6—Can We Improve on the EHW Variance Estimator?: The difference between $E[\widetilde{V}^{\text{ehw}}|N_1, N_0]$ (or the Neyman variance) and the total variance is equal to S_θ^2/n.

The term S_θ^2 is difficult to estimate because it depends on the unobserved differences $Y_i^*(1) - Y_i^*(0)$. As a result, S_θ^2/n is typically ignored in analyses of randomized experiments (see Imbens and Rubin (2015)). In particular, the EHW variance estimator implicitly sets the estimator of S_θ^2 to be equal to zero, resulting in conservative inference. For the case of a randomized experiment with a binary treatment, Aronow, Green, and Lee (2014) provided a lower bound for S_θ^2 based on the Fréchet–Hoeffding inequality. In Section 3, we propose an improved variance estimator that exploits the presence of fixed attributes.

The Appendix contains a Bayesian version of the analysis of the example from this section. Similarly to the results in this section, we show that, when the estimand of interest is defined for a finite population, the posterior variance depends not only on the sample sizes for treated and non-treated, but also on the respective population sizes. Also similarly to the analysis of this section, the posterior variance formula depends on whether the estimand is descriptive or causal.

3. THE GENERAL CASE

This section contains the main formal results in the article. We focus on a regression setting where we estimate a linear regression function for a scalar outcome and a number of regressors. The setting we consider here allows for the presence of two types of regressors: first, regressors that are causal, in the sense that they generate potential outcomes; and second, regressors that are attributes, in the sense that they are kept fixed for each unit in the thought experiment that provides the basis for inference. Which regressors are viewed as causal and which are viewed as attributes depends on the interpretation we wish to give to the regression estimates. If we wish to give a coefficient a causal interpretation, the corresponding regressor must be a cause. If a regressor is an attribute, the corresponding coefficient is simply estimating a population difference between subpopulations of units. For example, if we regress earnings on years of education, years of education may be the causal variable of interest. On the other hand, if we regress earnings on an indicator for participation in a job search program, age, and years of education, then the indicator for the program participation may be viewed as the causal variable of interest and age and years of education may be viewed as attributes. Because the repeated sampling thought experiment treats causes differently from attributes, the variance of the regression estimator will depend on this designation.

3.1. *Setup*

Consider a sequence of finite populations indexed by population size, n. Unit i in population n is characterized by a set of fixed attributes $Z_{n,i}$ (including an intercept) and by a potential outcome function, $Y_{n,i}^*(\cdot)$, which maps causes, $U_{n,i}$, into outcomes, with the realized outcome denoted by $Y_{n,i} = Y_{n,i}^*(U_{n,i})$. $Z_{n,i}$ and $U_{n,i}$ are real-valued column vectors, and $Y_{n,i}$ is scalar. We do not place restrictions on the types of the variables: they can be continuous, discrete, or mixed. Probabilities and expectations will be understood to be conditional on attributes and potential outcome functions, which remain fixed in our repeated sampling thought experiments. The realized outcomes in the sample vary from sample to sample because the units in the sample and the values of the causal variables change.

There is a sequence of samples associated with the sequence of populations. We will use $R_{n,i} = 1$ to indicate that unit i of population n is sampled, and $R_{n,i} = 0$ to indicate that it

is not sampled. For each unit in sample n, we observe the triple, $(Y_{n,i}, U_{n,i}, Z_{n,i})$. Relative to Section 2, we now allow for more complicated assignment mechanisms. In particular, we relax the assumption that the causes have identical distributions.

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

We assume independence of the treatment assignments. This is somewhat in contrast to the example in Section 2, where we fixed the marginal distribution of the regressor, allowing us to obtain exact finite sample results. We do not need this here because we are focused on asymptotic results. We can allow for some dependence in the assignment mechanism, for example, clustering of the type analyzed in Abadie, Athey, Imbens, and Wooldridge (2017).

For what follows, it is convenient to work with a transformation $X_{n,1}, \ldots, X_{n,n}$ of $U_{n,1}, \ldots, U_{n,n}$ that removes the correlation with the attributes. This can be accomplished in the following way. We assume that the population matrix $\sum_{i=1}^{n} Z_{n,i} Z'_{n,i}$ is full-rank. Then, define

$$X_{n,i} = U_{n,i} - \Lambda_n Z_{n,i}, \quad \text{where } \Lambda_n = \left(\sum_{i=1}^{n} E[U_{n,i}] Z'_{n,i}\right)\left(\sum_{i=1}^{n} Z_{n,i} Z'_{n,i}\right)^{-1}. \tag{3.1}$$

Later, we formally make an assumption that will guarantee that this transformation is well-defined for large n. It is important to notice that, because $\Lambda_n Z_{n,i}$ is deterministic in our setting and $U_{n,1}, \ldots, U_{n,n}$ are i.n.i.d., the variables $X_{n,1}, \ldots, X_{n,n}$ are i.n.i.d., too.

For population n, let $\mathbf{Y}_n$, $\mathbf{X}_n$, $\mathbf{Z}_n$, $\mathbf{R}_n$, and $\mathbf{Y}_n^*(\cdot)$ be matrices that collect outcomes, causes, attributes, sampling indicators, and potential outcome functions. We analyze the properties of the estimator $\widehat{\theta}_n$ obtained by minimizing least square errors in the sample:

$$(\widehat{\theta}_n, \widehat{\gamma}_n) = \underset{(\theta,\gamma)}{\arg\min} \sum_{i=1}^{n} R_{n,i}\big(Y_{n,i} - X'_{n,i}\theta - Z'_{n,i}\gamma\big)^2. \tag{3.2}$$

The properties of the population regression residuals, $e_{n,i} = Y_{n,i} - X'_{n,i}\theta_n - Z'_{n,i}\gamma_n$, depend on the exact nature of the estimands, (θ_n, γ_n). In what follows, we will consider alternative target parameters, which in turn will imply different properties for $e_{n,i}$. Notice also that, although the transformation in (3.1) is typically unfeasible (because the values of $E[U_{n,i}]$ may not be known), $\widehat{\theta}_n$ is not affected by the transformation in the sense that the least squares estimators $(\widetilde{\theta}_n, \widetilde{\gamma}_n)$, defined as

$$(\widetilde{\theta}_n, \widetilde{\gamma}_n) = \underset{(\theta,\gamma)}{\arg\min} \sum_{i=1}^{n} R_{n,i}\big(Y_{n,i} - U'_{n,i}\theta - Z'_{n,i}\gamma\big)^2,$$

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

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

ASSUMPTION 4—Random Sampling: (i) *There is a sequence of sampling probabilities*, ρ_n, *such that*

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

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

The first part of Assumption 4(ii) guarantees that as the population size increases, the (expected) sample size also increases. The second part of Assumption 4(ii) allows for the possibility that, as n increases, the sample size becomes a negligible fraction of the population size so that the EHW results, corresponding to $\rho = 0$, are included as a special case of our results.

The next assumption is a regularity condition bounding moments.

ASSUMPTION 5—Moments: *There exists some* $\delta > 0$ *such that the sequences*

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

are uniformly bounded.

Let

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

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

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

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

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

with analogous partitions for Ω_n, $\widetilde{W}_n$, and $\widetilde{\Omega}_n$. Notice that the transformation in (3.1) implies that Ω_n^{XZ} and Ω_n^{ZX} are matrices with all zero entries.

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

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

See the Appendix for proofs.

The next assumption imposes (deterministic) convergence of the expected value of the second moments in the population.

ASSUMPTION 6—Convergence of Moments: $\Omega_n \to \Omega$, *which is full rank*.

3.2. *Descriptive and Causal Estimands*

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

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

(i) Estimands are functionals of $(\mathbf{Y}_n^*(\cdot), \mathbf{X}_n, \mathbf{Z}_n, \mathbf{R}_n)$, permutation-invariant in the rows of the arguments.

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

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

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

We define three estimands of interest, which, under the conditions above, exist with probability approaching 1:

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

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

and

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

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

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

for the descriptive, causal-sample, and causal estimands, respectively. We will study the properties of the least squares estimator, $\widehat{\theta}_n$, defined by

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

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

Notice that $\theta_n^{\text{causal,sample}}$ and θ_n^{causal} are causal estimands, while θ_n^{descr} is not. However, the fact that an estimand is causal according to our definition does not imply it has an interpretation as an average causal effect. In Section 3.3, we present conditions under which the regression estimand does have such an interpretation.

3.3. *Causal Interpretations of the Estimands*

By construction, the descriptive estimand can be interpreted as the set of coefficients of a population best linear predictor (least squares) (e.g., Goldberger (1991)). A more challenging question concerns the interpretation of the two causal estimands, and in particular, their relation to the potential outcome functions. In this section, we investigate this question.

The next assumption is a generalization of random assignment and allows for a form of dependence of the assignment $U_{n,i}$ on the attributes $Z_{n,i}$.

ASSUMPTION 7—Expected Assignment: (i) *There exists a sequence of functions* h_n *such that*

$$E[U_{n,i}] = h_n(Z_{n,i}),$$

and (ii) *there exists a sequence of matrices* B_n *such that, for all* z,

$$h_n(z) = B_n z,$$

for all n large enough.

Assumption 7 looks very different from conventional exogeneity or unconfoundedness conditions, where the residuals are assumed to be (mean-) independent of the regressors, and so it merits some discussion. First, note that if the treatment $U_{n,i}$ is randomly assigned, $E[U_{n,i}]$ is constant and Assumptions 7(i) and (ii) are automatically satisfied as long as $Z_{n,i}$ includes an intercept.

Formally, Assumption 7 relaxes the completely randomized assignment setting by allowing the distribution of $U_{n,i}$ to depend on the attributes. However, this dependence is restricted in that the mean of $U_{n,i}$ is linear in $Z_{n,i}$. For example, Assumption 7 holds automatically when $U_{n,1}, \ldots, U_{n,n}$ are identically distributed and $Z_{n,i}$ contains a saturated set of indicators for all possible values of the attributes.

In the special case where the treatment is binary and $E[U_{n,i}]$ is the propensity score (Rosenbaum and Rubin (1983)), the assumption amounts to combination of an unconfoundedness assumption that the treatment assignment does not depend on the potential outcomes and a linear model for the propensity score.

Later in this section, we will show that under a set of conditions that includes Assumption 7, the two estimands θ_n^{causal} and $\theta^{\text{causal,sample}}$ can be interpreted as weighted averages of unit-level causal effects. The connection between linearity in the propensity score, represented in our analysis by $E[U_{n,i}] = B_n Z_{n,i}$, and the interpretation of population regression coefficients as weighted averages of heterogeneous causal effects has been previously noticed in related contexts (see Angrist (1998), Angrist and Pischke(2008), Aronow and Samii (2016), Słoczyński (2018)).

ASSUMPTION 8—Linearity of Potential Outcomes: *For all* u,

$$Y_{n,i}^*(u) = u'\theta_{n,i} + \xi_{n,i}, \tag{3.6}$$

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

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

The linearity in Assumption 8 is a strong restriction in many settings. However, in some leading cases—in particular, when the causal variable is binary or, more generally, when the causal variable takes on only a finite number of values—one can ensure that this assumption holds by including in $U_{n,i}$ indicator variables representing each but one of the possible values of the cause. With Assumption 8, we are able to provide a more transparent interpretation of the regression estimator.

THEOREM 1: *Suppose Assumptions 3–8 hold. Then, for all n large enough,*

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

and, with probability approaching 1,

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

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

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

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

and, with probability approaching 1,

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

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

COMMENT 7: Here, we provide a simple example that shows how the result in Theorems 1 and 2 may not hold in the absence of Assumption 7. Consider the population with three units described in Table III. For simplicity, we drop the subscript n. Notice that we could add replicates of these observations to make the example hold for any population size. In this example, $E[U_i] = 3bZ_i^2 - 2b$ is a nonlinear function of Z_i. Notice that

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

TABLE III
AN ARTIFICIAL EXAMPLE

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

so that $X_i = U_i$. Therefore, $E[X_i^2] = E[U_i^2]$. Also, because potential outcomes do not depend on X_i, it follows that $E[X_i Y_i] = E[X_i]Y_i^*(1) = E[U_i]Y_i^*(1)$. As a result,

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

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

3.4. *The Asymptotic Distribution of the Least Squares Estimator*

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

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

$$\varepsilon_{n,i} = Y_{n,i} - X'_{n,i}\theta_n^{\text{causal}} - Z'_{n,i}\gamma_n^{\text{causal}}. \tag{3.7}$$

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

COMMENT 9: We define the residuals here with respect to the population causal parameters θ_n^{causal} and γ_n^{causal}. Because we focus here on asymptotic results, the difference between the causal and descriptive parameters vanishes, and so defining the residuals in terms of the descriptive parameters would lead to the same results.

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

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

and the expected outer product

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

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

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

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

ASSUMPTION 9—Existence of Limits: Δ^{cond} *and* Δ^{ehw} *exist and are positive definite*.

THEOREM 3: *Suppose Assumptions 3–9 hold, and let* $\Gamma = \Omega^{XX} = \lim_{n\to\infty} \Omega_n^{XX}$. *Then,*

(i)

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

(ii)

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

(iii)

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

COMMENT 10: For both the population causal and the descriptive estimand, the asymptotic variance in the case with $\rho = 0$ reduces to the standard EHW variance, $\Gamma^{-1}\Delta^{\text{ehw}}\Gamma^{-1}$. If the sample size is non-negligible as a fraction of the population size, $\rho > 0$, the difference between the EHW variance and the finite population causal variance is positive semidefinite and equal to $\rho\Gamma^{-1}(\Delta^{\text{ehw}} - \Delta^{\text{cond}})\Gamma^{-1}$.

COMMENT 11—The Case With $\rho = 0$: The standard setting where we have a random sample from a large population is covered by the result in Theorem 3, part (i) or part (iii) with $\rho = 0$. For example, when we analyze data from the CPS or PSID, this seems a reasonable perspective. Even if the sampling from the U.S. population is not completely random, it is approximately so, and the sample is certainly small relative to the population. In that case, we do not need to worry about whether we are interested in a causal estimand because the standard methods are valid.

COMMENT 12—The Case With $\rho = 1$: The case where we observe all units in the population of interest, covered by the result in Theorem 3, part (i) with $\rho = 1$, or part (ii), is also common. For example, we may have all the states in the United States, or all the countries in the world, or all the individuals in the population of interest. In that case, taking account of the causal nature of the estimand is important, because a descriptive perspective would suggest the standard errors should be zero. This covers the case discussed in Manski and Pepper (2018).

COMMENT 13—A Convenience Sample: The setting where we have a convenience sample, where the relationship between the sample and the population is murky, is more complicated. For example, we may have all internet searches during a particular day, or all shopping trips to a single supermarket for a given week. This is, in our view, an important and common setting. In such settings, researchers often analyze the data, and report standard errors based on the sampling perspective, as if the sample is a random sample from a large population. Typically, they do so implicitly, by simply using standard methods without explicitly describing a sampling process. It seems a stretch to view the sample of shopping trips to a particular supermarket on a particular day as a random sample from the population of interest. At best, it is a systematic sample from the population of interest, for example, all individuals going to that particular supermarket, rather than a random sample from the population of individuals. However, there is no way to quantify the uncertainty arising from that sampling scheme without data from other periods or other supermarkets. In that case, we recommend to analyze the uncertainty relative to the causal sample estimand, and to be clear about what that estimand is in order to provide a conceptually precise measure of uncertainty.

COMMENT 14—The Case With $\rho \in (0, 1)$: While there are clearly many cases in practice where we do observe the entire population of interest, there are also settings where the following three things hold, at least approximately: (a) the population of interest is finite, (b) the sample is a random sample from this population, and (c) the ratio of sample size to population size ρ is known and large enough for this to matter. For example, Muralidharan and Niehaus (2017) discussed a number of randomized experiments in development economics where the study sample was drawn randomly from the population of interest. Keels et al. (2005) discussed using a 50% random sample in a mobility study, rather than the full population, for cost or computational reasons. The Integrated Public Use Microdata Series (IPUMS) data include a random sample of 10% of the census. Some other recent papers include DellaVigna, Lindner, Reizer, and Schmieder (2017), whose sample consists of a 50% de facto random sample of Hungarian citizens older than 14 and younger than 75 in 2002, Einav, Finkelstein, and Schrimpf (2015), who used a 20% random sample of Medicare Part D beneficiaries from 2007 to 2009, Hanna, Mullainathan, and Schwartzstein (2014), who randomly selected 117 (from the set of respondents) to participate in an experimental trial, and Farber (2015), who used a random subsample of 2/15 of the drivers in his data set. Another interesting case, with a more complex sample, is Munnell, Tootell, Browne, and McEneaney (1996), who used the population of mortgage applications in the city of Boston in 1990 for Black and Hispanic applicants (1200 obs) and a random sample of applications by White applicants (3300 obs) for the same city and year.

COMMENT 15: Presenting a general variance formula that includes $\rho = 0$ and $\rho = 1$ as special cases is helpful because it explicitly connects the two leading perspectives to uncertainty, sampling-based and design-based. It shows that there is no conceptual conflict between our proposed causal perspective and the standard sampling-based perspective on uncertainty, that our perspective merely adds a second source of uncertainty. It also shows that this perspective is particularly relevant when the researcher has observations on the entire population, a case that previously had not been satisfactorily addressed in the literature.

3.5. *The Variance Under Correct Specification*

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

ASSUMPTION 10—Constant Treatment Effects:

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

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

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

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

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

we obtain that equation (3.7) holds for $\gamma_n^{\text{causal}} = \Lambda_n'\theta_n + \lambda_n$ and $\varepsilon_{n,i} = \xi_{n,i} - Z_{n,i}'\lambda_n$. In this case, the residuals, $\varepsilon_{n,i}$, are non-stochastic. As a result, $E[X_{n,i}\varepsilon_{n,i}] = E[X_{n,i}]\varepsilon_{n,i} = 0$, which implies $\Delta^{\mu} = \Delta^{\text{ehw}} - \Delta^{\text{cond}} = 0$. This leads to the following result.

THEOREM 4: *Suppose that Assumptions 3–10 hold. Then*,

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

irrespective of the value of ρ.

Notice that the result of the theorem applies also with $\theta_n^{\text{causal,sample}}$ replacing θ_n^{causal} because the two parameter vectors are identical (with probability approaching 1) under Assumption 10.

COMMENT 16: The key insight in this theorem is that the asymptotic variance of $\widehat{\theta}_n$ does not depend on the ratio of the sample to the population size when the regression function is correctly specified. Therefore, it follows that the usual EHW variance matrix is correct for $\widehat{\theta}_n$ under these assumptions. For the special case with $X_{n,i}$ binary and no attributes beyond the intercept, this result can be inferred directly from Neyman's results for randomized experiments (Neyman (1923/1990)). In that case, the result of Theorem 4 follows from the restriction of constant treatment effects, $Y_{n,i}^*(1) - Y_{n,i}^*(0) = \theta_n$, which is extended to the more general case of non-binary regressors in Assumption 10. The asymptotic variance of $\widehat{\gamma}_n$, the least squares estimator of the coefficients on the attributes, still depends on the ratio of sample to population size, and it can be shown that the conventional robust EHW estimator continues to overestimate the variance of $\widehat{\gamma}_n$. For more details, see the earlier version of this paper (Abadie et al. (2014)).

4. ESTIMATING THE VARIANCE

We now turn to the problem of estimating the variance for the descriptive and causal estimands. In what follows, we will use the shorthands $V^{\text{causal}} = \Gamma^{-1}(\rho\Delta^{\text{cond}} +$

$(1-\rho)\Delta^{\text{ehw}})\Gamma^{-1}$, $V^{\text{causal,sample}} = \Gamma^{-1}\Delta^{\text{cond}}\Gamma^{-1}$, $V^{\text{descr}} = (1-\rho)\Gamma^{-1}\Delta^{\text{ehw}}\Gamma^{-1}$, and $V^{\text{ehw}} = \Gamma^{-1}\Delta^{\text{ehw}}\Gamma^{-1}$. There are four components to the asymptotic variances, ρ, Γ, Δ^{ehw}, and Δ^{cond}. The first three are straightforward to estimate. ρ can be estimated as $\widehat{\rho}_n = N/n$, as long as the population size is known. To estimate Γ, first estimate Λ_n as

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

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

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

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

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

The EHW large sample variance, V^{ehw}, is then estimated as

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

LEMMA 2: *Suppose Assumptions 3–7 and 9 hold with* $\delta = 4$. *Then*,

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

Let $\widehat{V}^{\text{descr}}_n = (1-\widehat{\rho}_n)\widehat{V}^{\text{ehw}}_n$. The result of Lemma 2 immediately implies $\widehat{V}^{\text{descr}}_n \xrightarrow{p} V^{\text{descr}}$.

It is more challenging to estimate V^{causal} and $V^{\text{causal,sample}}$ because they involve Δ^{cond}. Estimating Δ^{cond} is complicated because of the same reason that complicates the estimation of the variance of the average treatment effect estimator in Section 2. In that case, there are three terms in the expression for the variance in equation (2.3). The first two are straightforward to estimate, but the third one, S^2_θ/n, cannot be estimated consistently because we do not observe both potential outcomes for the same units. Often, researchers use the conservative estimator based on ignoring S^2_θ/n. If we proceed in the same fashion for the regression context of Section 3, we obtain the conservative estimator $\widehat{V}^{\text{ehw}}$, based on ignoring Δ^{μ}. We show, however, that in the presence of attributes, we can improve the variance estimator. We build on Abadie and Imbens (2008), Abadie, Imbens, and Zheng (2014), and Fogarty (2016) who, in contexts different than the one studied in this article, have used the explanatory power of attributes to improve variance estimators. Abadie and Imbens (2008) and Abadie, Imbens, and Zheng (2014) did so using nearest-neighbor techniques. Here, we follow Fogarty (2016) and apply linear regression techniques. The proposed estimator replaces the expectations $E[X_{n,i}\varepsilon_{n,i}]$, which cannot be consistently estimated, with predictors from a linear least squares projection of estimates of $X_{n,i}\varepsilon_{n,i}$ on the attributes, $Z_{n,i}$. Let $\widehat{X}_{n,i} = U_{n,i} - \widehat{\Lambda}_n Z_{n,i}$, and

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

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

ASSUMPTION 11:

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

has a limit.

Consider now the following estimator:

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

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

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

Estimators of $V^{\text{causal,sample}}$ and V^{causal} follow immediately from Lemma 3 by replacing Δ^{cond} with the estimate $\widehat{\Delta}_n^Z$ in the asymptotic variance formulas of Theorem 3, leading to $\widehat{V}_n^{\text{causal,sample}} = \widehat{\Gamma}_n^{-1}\widehat{\Delta}_n^Z\widehat{\Gamma}_n^{-1}$ for the estimation of $V^{\text{causal,sample}}$ and $\widehat{V}_n^{\text{causal}} = \widehat{\rho}_n\widehat{V}_n^{\text{causal,sample}} + (1-\widehat{\rho}_n)\widehat{V}_n^{\text{ehw}}$ for the estimation of V^{causal}. These estimators are not larger (and typically smaller) than $\widehat{V}_n^{\text{ehw}}$ and they remain conservative in large samples.

COMMENT 17: A special case of the adjusted variance arises when $Z_{n,i}$ is a set of exhaustive and mutually exclusive dummy variables, or if we reduce the information in $Z_{n,i}$ to such indicators. Then, the residuals from regressing $\widehat{X}_{n,i}\widehat{\varepsilon}_{n,i}$ on $Z_{n,i}$ are simply stratum-specific demeaned versions of $\widehat{X}_{n,i}\widehat{\varepsilon}_{n,i}$, and a conservative estimator of Δ^{cond} can be obtained using the variance formulas in Wooldridge (2001) for standard stratified samples.

5. SIMULATIONS

In this section, we use a simple data-generating process as well as simulations to illustrate the difference between the conventional EHW variance estimator and the variance estimators proposed in this article. We focus on the case of a single causal variable, $X_{n,i}$. In addition to the causal variable, the simulations employ an outcome variable, $Y_{n,i}$, and a vector of attributes, $Z_{n,i}$, which consists of a constant equal to 1 and k values drawn independently from the standard normal distribution. The potential outcome function has the form in equation (3.6). Population values of $\theta_{n,i}$ are generated as independent draws from a normal distribution with mean $Z'_{n,i}\psi$, where $\psi = (0, \psi_1, \ldots, \psi_k)'$ is a $k+1$ vector, and variance σ_θ^2. Population values of $\xi_{n,i}$ and $U_{n,i}$ are generated as independent draws

from a normal distribution with mean 0 and variance 1. Because in this data-generating process $E[U_{n,i}] = 0$, it follows that Λ_n is a row-vector of zeros and $X_{n,i} = U_{n,i}$. We use this data-generating process to produce a population of size n. For this data-generating process, it can be shown that $\Gamma = 1$, $\Delta^{\text{ehw}} = 1 + 3(\psi'\psi + \sigma_\theta^2)$ and $\Delta^{\text{cond}} = 1 + 2(\psi'\psi + \sigma_\theta^2)$, and $\Delta^Z = 1 + 2\psi'\psi + 3\sigma_\theta^2$ with probability 1. In each simulation repetition, we sample units at random with probability ρ from the population. As a result, the sample size N is random with $E[N] = n\rho$. For each sample, we estimate $\widehat{\theta}_n$ by least squares (as in equation (3.2)) and a number of variance estimators.

In Table IV, we report the results of the simulations. We consider seven designs. The first column reports the basic design, with $\rho = 0.01$ and $n = 100{,}000$, so the average sample size is 1000. In this design, there is one stochastic regressor, so $k = 1$, and the distribution of the treatment effect, $\theta_{n,i}$, is given by parameter values $\psi = (0, 2)'$ and $\sigma_\theta^2 = 1$. The remaining designs in the second to seventh columns are variations of the basic design in the first column. In the second design, we increase the dimensionality of $Z_{n,i}$ used for estimation from two to ten. Still, in this design ψ has all entries equal to zero except for $\psi_1 = 2$, so only the first stochastic regressor matters for the distribution of $\theta_{n,i}$. In the next design, we change the population size to 10,000, so that the average sample size is 100. In the fourth design, we change the population size to 1000 and the sampling rate, ρ, to 1. In the fifth design, we impose $\psi'\psi = 0$, which makes the treatment effect unrelated to the regressors, $Z_{n,i}$. In the sixth design, we set $\sigma_\theta^2 = 0$, which removes the stochastic part of the treatment effect. In the last design, $\psi'\psi = 0$ and $\sigma_\theta^2 = 0$, so the treatment effect is constant. The first panel of Table IV provides the parameters of each of the seven simulation designs.

The second panel of Table IV reports the standard deviations of $(\widehat{\theta}_n - \theta_n^{\text{descr}})$, $(\widehat{\theta}_n - \theta_n^{\text{causal,sample}})$, and $(\widehat{\theta}_n - \theta_n^{\text{causal}})$ across simulation iterations. The remaining panels report feasible standard errors based on the estimators of Section 4 as well as bootstrap standard errors, along with coverage rates of the corresponding 95 percent confidence intervals. We employ 50,000 iterations for the simulations and 1000 bootstrap samples. The coverage rates in each of the panels of the table are based on adding and subtracting 1.96 times the standard errors whose average appears in the first row of that panel.

For the basic design in the first column of Table IV, $\rho = 0.01$ is very small, and EHW and bootstrap standard errors provide accurate estimates of the standard deviations of $(\widehat{\theta}_n - \theta_n^{\text{descr}})$ and $(\widehat{\theta}_n - \theta_n^{\text{causal}})$. However, the standard deviation of $(\widehat{\theta}_n - \theta_n^{\text{causal,sample}})$ is substantially smaller than that of $(\widehat{\theta}_n - \theta_n^{\text{descr}})$ and $(\widehat{\theta}_n - \theta_n^{\text{causal}})$, and the EHW and bootstrap variance estimators are very conservative for the sample average causal effect, $\theta_n^{\text{causal,sample}}$. The variance estimator based on $\widehat{V}_n^{\text{causal,sample}}$ is substantially smaller, and still has more than correct coverage for $\theta^{\text{causal,sample}}$. Increasing the number of regressors in the second design leaves the result virtually unaffected. The same patterns of results appear in the third column, albeit with less precise variance estimators due to much smaller sample sizes. In the fourth design, $\rho = 1$ and, as predicted by the results in Section 4, EHW standard errors greatly overestimate the variability of $(\widehat{\theta}_n - \theta_n^{\text{descr}})$. The same is true for bootstrap standard errors. In the fifth design, we go back to the small sampling rate, $\rho = 0.01$ and this time $\psi'\psi = 0$, so regressors do not explain variation in treatment effects, and $\Delta^Z = \Delta^{\text{ehw}}$. As suggested by the results in Sections 3 and 4, all variance estimators produce similar results in this design. In the sixth design, where regressors explain all the variation in treatment effects, $\Delta^Z = \Delta^{\text{cond}}$ and standard errors based on $\widehat{V}_n^{\text{causal,sample}}$ and $\widehat{V}_n^{\text{causal}}$ closely approximate the standard deviations of $(\widehat{\theta}_n - \theta_n^{\text{causal,sample}})$ and $(\widehat{\theta}_n - \theta_n^{\text{causal}})$, respectively. In the final design with a constant treatment effect, all the variances are similar.

TABLE IV

SIMULATION RESULTS WITH COVERAGE FOR NOMINAL 95% CONFIDENCE INTERVALS

$E[N] = \rho n$	1000	1000	100	1000	1000	1000	1000
ρ	0.01	0.01	0.01	1	0.01	0.01	0.01
k	1	10	1	1	1	1	1
$\psi'\psi$	4	4	4	4	0	4	0
σ_θ^2	1	1	1	1	1	0	0
$\text{sd}(\widehat{\theta}_n - \theta_n^{\text{descr}})$	0.125	0.126	0.399	0.000	0.063	0.113	0.031
$\text{sd}\,\widehat{\theta}_n - \theta_n^{\text{causal,sample}})$	0.105	0.104	0.331	0.100	0.055	0.095	0.032
$\text{sd}\,\widehat{\theta}_n - \theta_n^{\text{causal}})$	0.125	0.126	0.400	0.100	0.063	0.114	0.032
Average $(\widehat{V}_n^{\text{ehw}}/N)^{1/2}$	0.125	0.124	0.370	0.121	0.063	0.113	0.032
Coverage θ_n^{descr}	0.949	0.947	0.923	1.000	0.948	0.947	0.950
Coverage $\theta_n^{\text{causal,sample}}$	0.980	0.981	0.969	0.982	0.974	0.981	0.950
Coverage θ_n^{causal}	0.948	0.947	0.922	0.982	0.947	0.947	0.950
Average $(\widehat{V}_n^{\text{boot}}/N)^{1/2}$	0.126	0.127	0.426	0.122	0.064	0.115	0.032
Coverage θ_n^{descr}	0.950	0.950	0.955	1.000	0.950	0.949	0.953
Coverage $\theta_n^{\text{causal,sample}}$	0.981	0.982	0.986	0.982	0.975	0.981	0.951
Coverage θ_n^{causal}	0.950	0.949	0.955	0.982	0.949	0.949	0.951
Average $(\widehat{V}_n^{\text{desc}}/N)^{1/2}$	0.124	0.124	0.368	0.000	0.063	0.113	0.031
Coverage θ_n^{descr}	0.948	0.946	0.921	1.000	0.947	0.946	0.949
Coverage $\theta_n^{\text{causal,sample}}$	0.980	0.980	0.968	0.000	0.973	0.981	0.948
Coverage θ_n^{causal}	0.947	0.946	0.921	0.000	0.946	0.946	0.948
Average $(\widehat{V}_n^{\text{causal,sample}}/N)^{1/2}$	0.108	0.107	0.317	0.104	0.063	0.094	0.032
Coverage θ_n^{descr}	0.908	0.905	0.872	1.000	0.948	0.894	0.950
Coverage $\theta_n^{\text{causal,sample}}$	0.956	0.957	0.937	0.957	0.974	0.948	0.949
Coverage θ_n^{causal}	0.907	0.904	0.870	0.957	0.947	0.892	0.949
Average $(\widehat{V}_n^{\text{causal}}/N)^{1/2}$	0.125	0.124	0.369	0.104	0.063	0.113	0.032
Coverage θ_n^{descr}	0.949	0.947	0.922	1.000	0.948	0.947	0.950
Coverage $\theta_n^{\text{causal,sample}}$	0.980	0.981	0.969	0.957	0.974	0.981	0.950
Coverage θ_n^{causal}	0.948	0.947	0.922	0.957	0.947	0.946	0.950

6. CONCLUSION

In this article, we study the interpretation of standard errors in regression analysis when the assumption that the sample is drawn randomly from a much larger population of interest is not appropriate. We base our results on a potential outcome framework, where the estimands of interest may be descriptive or causal, and we provide a coherent interpretation for standard errors that allows for uncertainty coming from both random sampling and from conditional random assignment. The standard errors estimators proposed in this article may be different from the conventional ones, and they may vary depending on (i) the specific nature of the estimand of interest (i.e., descriptive or causal), (ii) the fraction of the population represented in the sample, and (iii) the extent to which measured attributes explain variation in treatment effects.

In the current article, we focus exclusively on linear regression models. The concerns we raise in this article arise in many other settings and for other kinds of hypotheses, and the implications would need to be worked out for those settings. Thus, we see this article as a first step in a broader research program.

APPENDIX

A.1. *A Bayesian Approach*

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

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

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

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

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

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

$$\begin{pmatrix} Y_i^*(1) \\ Y_i^*(0) \end{pmatrix} \Big| \mu_1, \mu_0, \sigma_1^2, \sigma_0^2, \kappa \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \kappa\sigma_1\sigma_0 \\ \kappa\sigma_1\sigma_0 & \sigma_0^2 \end{pmatrix} \right),$$

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

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

The descriptive estimand is

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

The causal estimand is

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

To summarize, if the population is large, relative to the sample, the posterior distributions of θ_n^{descr}, θ_n^{causal}, and θ^{causal} agree. However, if the population is small, the three posterior distributions differ, and the researcher needs to be precise in defining the estimand.

In such cases, simply using the posterior of the super-population estimand $\theta^{\text{causal}} = \mu_1 - \mu_0$ may not be appropriate, because the posterior inferences for θ^{causal} will differ from those of θ_n^{causal} or θ_n^{descr}.

A.2. *Proofs*

PROOF OF LEMMA 1: See Supplemental Material. *Q.E.D.*

PROOF OF THEOREM 1: For n large enough, $\sum_{i=1}^n Z_{n,i} Z_{n,i}'$ is full rank and Λ_n exists, so $\Omega_n^{ZX} = 0$. This implies

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

Moreover, for n large enough, $\Lambda_n = B_n$, which implies $E[X_{n,i}] = 0$, $\widetilde{\Omega}_n^{XZ} = 0$, and

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

with probability approaching 1. Now,

$$\begin{aligned} E[X_{n,i} Y_{n,i}] &= E\big[X_{n,i} U_{n,i}'\big]\theta_{n,i} + E[X_{n,i}]\xi_{n,i} \\ &= E\big[X_{n,i} X_{n,i}'\big]\theta_{n,i} \end{aligned}$$

implies the results. *Q.E.D.*

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

The following lemma will be useful for establishing asymptotic normality.

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

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

is bounded for some $\delta > 0$,

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

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

and

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

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

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

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

PROOF: Notice that

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

and

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

Now the continuous mapping theorem implies

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

As a result, it is enough to prove

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

Let

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

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

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

and

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

Therefore,

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

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

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

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

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

Now, the result of the lemma follows from equation (A.1) and from

$$s_n/\sqrt{\sigma^2 + (1-\rho)\kappa^2} \to 1. \qquad \textit{Q.E.D.}$$

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

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

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

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

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

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

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

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

Let $a \neq 0$. Then,

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

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

This implies

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

Using the Cramer–Wold device, this implies

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

The proofs of (ii) and (iii) are similar. *Q.E.D.*

PROOF OF THEOREM 3: To prove (i), notice that

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

is invertible with probability approaching 1. Then,

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

Therefore,

$$\begin{aligned}\sqrt{N}\begin{pmatrix}\widehat{\theta}_n - \theta_n^{\text{causal}} \\ \widehat{\gamma}_n - \gamma_n^{\text{causal}}\end{pmatrix} &= \left(\frac{1}{N}\sum_{i=1}^{n}R_{n,i}\begin{pmatrix} X_{n,i}X_{n,i}' & X_{n,i}Z_{n,i}' \\ Z_{n,i}X_{n,i}' & Z_{n,i}Z_{n,i}' \end{pmatrix}\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{n}R_{n,i}\begin{pmatrix} X_{n,i}\varepsilon_{n,i} \\ Z_{n,i}\varepsilon_{n,i}\end{pmatrix} \\ &= \begin{pmatrix}\Omega_n^{XX} & \Omega_n^{XZ} \\ \Omega_n^{ZX} & \Omega_n^{ZZ}\end{pmatrix}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{n}R_{n,i}\begin{pmatrix} X_{n,i}\varepsilon_{n,i} \\ Z_{n,i}\varepsilon_{n,i}\end{pmatrix} + r_n,\end{aligned}$$

where

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

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

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

if we can show

$$(1/\sqrt{N})\sum_{i=1}^n R_{n,i}Z_{n,i}\varepsilon_{n,i} = O_p(1).$$

We can write this standardized sum as

$$(n\rho_n/N)^{1/2}\left[n^{-1/2}\sum_{i=1}^n (R_{n,i}/\sqrt{\rho_n})Z_{n,i}\varepsilon_{n,i}\right].$$

As shown in Lemma A.1, $(n\rho_n/N)^{1/2} \xrightarrow{p} 1$. Therefore, it suffices to show

$$n^{-1/2}\sum_{i=1}^n (R_{n,i}/\sqrt{\rho_n})Z_{n,i}\varepsilon_{n,i} = O_p(1).$$

This expression has zero mean because $R_{n,i}$ is independent of $\varepsilon_{n,i}$ and

$$\sum_{i=1}^n Z_{n,i}\mathrm{E}(\varepsilon_{n,i}) = 0.$$

We can study each element of the vector separately. By Chebyshev's inequality, it suffices to show that the variances are bounded. Consider the jth element. Then, by independence across i,

$$\begin{aligned}\operatorname{var}\left[n^{-1/2}\sum_{i=1}^n (R_{n,i}/\sqrt{\rho_n})Z_{n,i,j}\varepsilon_{n,i}\right] &= n^{-1}\sum_{i=1}^n \operatorname{var}\big[(R_{n,i}/\sqrt{\rho_n})Z_{n,i,j}\varepsilon_{n,i}\big] \\ &\le n^{-1}\sum_{i=1}^n E\big\{\big[(R_{n,i}/\sqrt{\rho_n})Z_{n,i,j}\varepsilon_{n,i}\big]\big\}^2 \\ &= n^{-1}\sum_{i=1}^n Z_{n,i,j}^2 E\big(\varepsilon_{n,i,j}^2\big),\end{aligned}$$

where the last equality holds because $E[R_{n,i}] = \rho_n$ and $Z_{n,i,j}$ is non-random. Both $Z_{n,i,j}^2$ and $E[\varepsilon_{n,i}^2]$ are bounded by Assumption 5, and so this completes the proof. The proofs of (ii) and (iii) are analogous. *Q.E.D.*

PROOF OF THEOREM 4: The result follows directly from $E[X_{n,i}\varepsilon_{n,i}] = 0$. *Q.E.D.*

PROOF OF LEMMA 2: First, notice that (with probability approaching 1) Λ_n exists and it is equal to B_n. This implies

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

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

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

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

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

and

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

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

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

and

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

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

PROOF OF LEMMA 3: Notice that

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

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

Let

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

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

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

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

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

Let

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

Notice that

$$\begin{aligned}\Delta_n^{\mu} - \Delta_n^{\text{proj}} = {} & \frac{1}{n}\sum_{i=1}^{n} E[X_{n,i}\varepsilon_{n,i}]E\big[\varepsilon_{n,i}X'_{n,i}\big] \\ & - \left(\frac{1}{n}\sum_{i=1}^{n} E[X_{n,i}\varepsilon_{n,i}]Z'_{n,i}\right)\left(\frac{1}{n}\sum_{i=1}^{n} Z_{n,i}Z'_{n,i}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{n,i}E\big[\varepsilon_{n,i}X'_{n,i}\big]\right).\end{aligned}$$

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

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

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

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

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

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

where Δ^{cond}, Δ^Z, and Δ^{ehw} are the limits of Δ_n^{cond}, Δ_n^Z, and Δ_n^{ehw}, respectively. *Q.E.D.*

REFERENCES

ABADIE, A., S. ATHEY, G. W. IMBENS, AND J. M. WOOLDRIDGE (2014): "Finite Population Causal Standard Errors," Technical Report, National Bureau of Economic Research. [265,281]

——— (2017): "When Should You Adjust Standard Errors for Clustering?" Technical Report, National Bureau of Economic Research. [273]

——— (2020): "Supplement to 'Sampling-Based versus Design-Based Uncertainty in Regression Analysis'," *Econometrica Supplemental Material*, 88, https://doi.org/10.3982/ECTA12675. [268]

ABADIE, A., AND G. W. IMBENS (2008): "Estimation of the Conditional Variance in Paired Experiments," *Annales d'Economie et de Statistique*, 91/92, 175–187. [282]

ABADIE, A., G. W. IMBENS, AND F. ZHENG (2014): "Inference for Misspecified Models With Fixed Regressors," *Journal of the American Statistical Association*, 109 (508), 1601–1614. [282]

ANGRIST, J., AND S. PISCHKE (2008): *Mostly Harmless Econometrics: An Empiricists' Companion*. Princeton, NJ: Princeton University Press. [276]

ANGRIST, J. D. (1998): "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66 (2), 249–288. [276]

ARONOW, P. M., D. P. GREEN, AND D. K. LEE (2014): "Sharp Bounds on the Variance in Randomized Experiments," *The Annals of Statistics*, 42 (3), 850–871. [272]

ARONOW, P. M., AND C. SAMII (2016): "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science*, 60 (1), 250–267. [276]

DAVIDSON, J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*. Advanced Texts in Econometrics. Oxford: Oxford University Press. [290]

DEATON, A. (2010): "Instruments, Randomization, and Learning About Development," *Journal of Economic Literature*, 48 (2), 424–455. [267,271]

DELLAVIGNA, S., A. LINDNER, B. REIZER, AND J. F. SCHMIEDER (2017): "Reference-Dependent Job Search: Evidence From Hungary," *The Quarterly Journal of Economics*, 132 (4), 1969–2018. [280]

DONOHUE, J. J., A. ANEJA, AND K. WEBER (2019): "Right-to-Carry Laws and Violent Crime: A Comprehensive Assessment Using Panel Data and a State-Level Synthetic Controls Analysis," *Journal of Empirical Legal Studies*, 16 (2), 198–247. [268]

EICKER, F. (1967): "Limit Theorems for Regressions With Unequal and Dependent Errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley, CA: University of California Press, 59–82. [267]

EINAV, L., A. FINKELSTEIN, AND P. SCHRIMPF (2015): "The Response of Drug Expenditure to Nonlinear Contract Design: Evidence From Medicare Part D," *The Quarterly Journal of Economics*, 130 (2), 841–899. [280]

FARBER, H. S. (2015): "Why You Can't Find a Taxi in the Rain and Other Labor Supply Lessons From Cab Drivers," *The Quarterly Journal of Economics*, 130 (4), 1975–2026. [280]

FOGARTY, C. B. (2016): "Regression Assisted Inference for the Average Treatment Effect in Paired Experiments," Preprint. Available at arXiv:1612.05179. [282]

FREEDMAN, D. (2008): "On Regression Adjustmens to Experimental Data," *Advances in Applied Mathematics*, 40 (2), 180–193. [266,267]

GOLDBERGER, A. S. (1991): *A Course in Econometrics*. Cambridge, MA: Harvard University Press. [276]

HANNA, R., S. MULLAINATHAN, AND J. SCHWARTZSTEIN (2014): "Learning Through Noticing: Theory and Evidence From a Field Experiment," *The Quarterly Journal of Economics*, 129 (3), 1311–1353. [280]

HOLLAND, P. W. (1986): "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81 (396), 945–960. [268]

HUBER, P. J. (1967): "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley, CA: University of California Press, 221–233. [267]

IMBENS, G. W., AND D. B. RUBIN (2015): *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press. [266,268,272]

KEELS, M., G. J. DUNCAN, S. DELUCA, R. MENDENHALL, AND J. ROSENBAUM (2005): "Fifteen Years Later: Can Residential Mobility Programs Provide a Long-Term Escape From Neighborhood Segregation, Crime, and Poverty," *Demography*, 42 (1), 51–73. [267,280]

LIN, W. (2013): "Agnostic Notes on Regression Adjustments for Experimental Data: Reexamining Freedman's Critique," *The Annals of Applied Statistics*, 7 (1), 295–318. [266,267]

LU, J. (2016): "On Randomization-Based and Regression-Based Inferences for 2^K Factorial Designs," *Statistics & Probability Letters*, 112, 72–78. [266]

MACKINNON, J. G., AND H. WHITE (1985): "Some Heteroskedasticity-Consistent Covariance Matrix Estimators With Improved Finite Sample Properties," *Journal of Econometrics*, 29 (3), 305–325. [271]

MANSKI, C. F. (2013): *Public Policy in an Uncertain World: Analysis and Decisions*. Cambridge, MA: Harvard University Press. [267,271]

MANSKI, C. F., AND J. V. PEPPER (2018): "How Do Right-to-Carry Laws Affect Crime Rates? Coping With Ambiguity Using Bounded-Variation Assumptions," *Review of Economics and Statistics*, 100 (2), 232–244. [266-268,279]

MUNNELL, A. H., G. M. TOOTELL, L. E. BROWNE, AND J. MCENEANEY (1996): "Mortgage Lending in Boston: Interpreting HMDA Data," *The American Economic Review*, 86 (1), 25–53. [280]
MURALIDHARAN, K., AND P. NIEHAUS (2017): "Experimentation at Scale," *Journal of Economic Perspectives*, 31 (4), 103–124. [267,280]
NEYMAN, J. (1923/1990): "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Statistical Science*, 5 (4), 465–472. [266-268,270,271,281]
ROSENBAUM, P. R. (2002): *Observational Studies*. Berlin: Springer. [266]
ROSENBAUM, P. R., AND D. B. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70 (1), 41–55. [276]
RUBIN, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66 (5), 688–701. [268]
SAMII, C., AND P. M. ARONOW (2012): "On Equivalencies Between Design-Based and Regression-Based Variance Estimators for Randomized Experiments," *Statistics & Probability Letters*, 82 (2), 365–370. [266, 267]
SHADISH, W. R., T. D. COOK, AND D. T. CAMPBELL (2002): *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton, Mifflin and Company. [267,271]
SŁOCZYŃSKI, T. (2018): "A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands," IZA Discussion Paper 11866. [276]
WHITE, H. (1980a): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48 (1), 817–838. [267]
——— (1980b): "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21 (1), 149–170. [267]
——— (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50 (1), 1–25. [267]
WOOLDRIDGE, J. M. (2001): "Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples," *Econometric Theory*, 17 (2), 451–470. [283]

Co-editor Ulrich K. Müller handled this manuscript.

Manuscript received 22 July, 2014; final version accepted 21 June, 2019; available online 31 July, 2019.