The Generalized Oaxaca-Blinder Estimator

Kevin Guo and Guillaume Basse

Department of Statistics, Stanford University, Stanford, CA

ABSTRACT
After performing a randomized experiment, researchers often use ordinary least-square (OLS) regression to adjust for baseline covariates when estimating the average treatment effect. It is widely known that the resulting confidence interval is valid even if the linear model is misspecified. In this article, we generalize that conclusion to covariate adjustment with nonlinear models. We introduce an intuitive way to use any "simple" nonlinear model to construct a covariate-adjusted confidence interval for the average treatment effect. The confidence interval derives its validity from randomization alone, and when nonlinear models fit the data better than linear models, it is narrower than the usual interval from OLS adjustment.

1. Introduction
1.1. Motivation
In this article, we study how covariates can be used to construct more precise estimates of the sample average treatment in a completely randomized experiment. Our investigation was inspired by Lin (2013), who suggested performing covariate adjustment by simply fitting an OLS model with treatment-by-covariate interactions.

\[ Y_i = \mu + \tau Z_i + \beta^\top (x_i - \bar{x}) + \gamma^\top Z_i (x_i - \bar{x}) \]  \hspace{1cm} (1)

In the above display, \( Y_i \in \mathbb{R} \) is an outcome variable, \( Z_i \in \{0, 1\} \) denotes the treatment status of unit \( i \) (1 for "treatment", 0 for "control"), and \( x_i \in \mathbb{R}^d \) is a vector of baseline covariates. Remarkably, Lin showed that even if the model (1) is arbitrarily misspecified, random assignment of \( Z_i \) is enough to justify standard inferences based on the regression coefficient \( \hat{\beta} \). This result is summarized (informally) in Theorem 1.

Theorem 1. (Lin's result, informal)
Let \( n_1 \) be the size of the treatment group, and let \( n_0 \) be the size of the control group. If \( \min(n_0, n_1) \gg d \), then we have

\[ \sqrt{n}(\hat{\tau} - \tau) \sim N(0, \sigma^2) \]  \hspace{1cm} (2)

where \( \tau \) is the sample average treatment effect. Moreover, \( \hat{\tau} \) is at least as efficient as Neyman’s (1923) unadjusted difference-of-means estimator, and the usual confidence interval for \( \tau \) based on Huber–White "robust" standard errors is valid.

This result initially came as a major surprise to the causal inference community. Although cosmetically similar results appear in earlier work by Yang and Tsiatis (2001), their proofs rely on the assumption that experimental units are iid samples from a larger population. Thus, they do not apply to the convenience samples commonly encountered in social science experiments and clinical trials (Olsen et al. 2013; Rosenberger, Ushner, and Wang 2019; Abadie et al. 2020). Indeed, before Lin’s work, it was widely believed that regression adjustment to experimental data could not be justified on the basis of randomization alone (Freedman 2008a, b, c; Schochet 2010). The accepted view was summarized in the abstract of Freedman (2008a): "since randomization does not justify the [regression] models, almost anything can happen." Lin’s result shows this is not the case.

Theorem 1 has since been generalized to other experimental designs (Fogarty 2018; Li and Ding 2020; Liu and Yang 2020) and to high-dimensional linear regression (Bloniarz et al. 2016; Lei and Ding 2020). It is now widely known that covariate adjustment with linear models is never "wrong." However, that does not mean it is always "right." For example, when the outcome variable is binary, nonnegative, or highly skewed, one suspects that it may be possible to further improve precision by using nonlinear models. There have been various clever proposals for how this might be done, but none of them have all four of the appealing properties of Lin’s result:

1. Statistic inference.
   The method produces a confidence interval with rigorous mathematical guarantees.

2. Robustness to misspecification.
   The method does not require any specific assumptions about the relationship between covariates and outcomes to be valid.

3. Randomization-based.
   The only probabilistic assumption is that treatment assignments \( Z_i \) are randomly assigned. Validity should not be compromised if the experimental units are not randomly sampled from a larger population.

4. Computational simplicity.
   The estimator can be computed by practitioners without extensive programming ability, using only functions that already exist in most statistical software packages.
Many proposals come close. Rosenbaum (2002) suggested forming a confidence interval by inverting a Fisher randomization test based on the residuals of an arbitrary (possibly nonlinear) model; this satisfies 1, 3, and (arguably) 4, but the validity of the confidence interval requires a constant additive treatment effect. The leave-one-out potential outcomes methods of Wu and Gagnon-Bartsch (2018) used any regression model to construct a randomization-unbiased estimate of $\tau$, but it is not simple to implement\(^1\) and does not come with a confidence interval. The literature on doubly robust methods (Robins, Rotnitzky, and Zhao 1994a; Kang and Schafer 2007; Cao, Tsatis, and Davidian 2009; Rosenblum and van der Laan 2011) is full of proposals satisfying 1, 2, and 4, but their theoretical justifications always assume that experimental units are iid samples from a larger population.

The purpose of this article is to introduce a general-purpose method for using any sufficiently "simple" nonlinear regression model to perform covariate adjustment in a manner that satisfies 1–4. Almost every widely used parametric model is simple enough to work, and so are some nonparametric models. As long as the chosen nonlinear models fit the data better than linear models, then our confidence intervals are narrower than the robust standard error confidence interval from OLS adjustment.

1.2. The Potential Outcomes Model

In this article, we use the Neyman–Rubin potential outcomes model of causality (Neyman 1923; Rubin 1974). We consider a finite population of $n$ experimental units, indexed by the set $I = \{1, 2, \ldots, n\}$. Each experimental unit consists of a triple $(y_{1i}, y_{0i}, x_i)$, where $x_i \in \mathbb{R}^d$ is a vector of covariates and $y_{1i}, y_{0i}$ are potential outcomes.$^2$ The goal is to estimate the sample average treatment effect $\tau = 1/n \sum_{i=1}^{n} (y_{1i} - y_{0i})$.

We adopt the framework of randomization inference, which treats all of the quantities $\{(y_{1i}, y_{0i}, x_i)^\tau\}_{i=1}^{n}$ as fixed constants. The only randomness is in the treatment assignments $(Z_1, \ldots, Z_n) \sim \mathbb{P}_{n,1}$, where $\mathbb{P}_{n,1}$ is the uniform distribution on length-$n$ binary vectors with $||v||_1 = n_1$. The observed outcome is $y_i = Z_i y_{1i} + (1 - Z_i) y_{0i}$.

This “fixed-population” framework was originally introduced by Neyman (1923), and it has been used extensively throughout statistics. In survey sampling, it is called “design-based inference,” and it is the predominant mode of analysis in that literature (Särndal, Swensson, and Wretman 2003; Lohr 2009). More recently, randomization inference has seen a revival in causal inference, both in terms of theoretical developments (Bloniarz et al. 2016; Li and Ding 2017; Mukerjee, Dasgupta, and Rubin 2018; Li and Ding 2020) and in terms of its popularity with applied researchers (Hansen and Bowers 2009; Athey and Imbens 2017). The manifold connections between the fixed-population frameworks in survey sampling and randomized experiments are explored in Fienberg et al. (2018) and Mukerjee, Dasgupta, and Rubin (2018).

Randomization inference is in contrast with “superpopulation” inference, which assumes that experimental units are themselves random samples from a hypothetical infinite population, $(y_{1i}, y_{0i}, x_i) \sim \mathbb{P}$. Although the superpopulation framework is mathematically convenient, it is unnatural in many settings. For example, in a randomized experiment where every county in a state is assigned to either “treatment” or “control,” it is not obvious what infinite “population” of counties the distribution $\mathbb{P}$ might refer to. In other experiments, experimental units are either convenience samples or they are carefully selected based on their observed covariates (e.g., to ensure representativeness). In both of these settings, the participants cannot be reasonably modeled as a random sample from any population. In these problems, inferences based on randomization alone are still meaningful, while inferences based on random sampling may be suspect. See Abadie et al. (2020) for a more thorough discussion.

2. Generalizing Oaxaca-Blinder

2.1. Beyond Linear Adjustment

In order to motivate our procedure, we first present another way of looking at Lin’s “interactions” estimator. Although the estimator is defined as a coefficient in a regression model, that characterization of $\hat{\tau}$ does not illuminate why it works. For example, it is not obvious from that characterization that covariates must be centered—including dummy variables coding categorical features—in order for $\hat{\tau}$ to have model-free validity. Without centering, the “interactions” estimator may be badly biased even in large samples.

The reason why centering is so important is that, if all covariates are centered, then fitting the model (1) is equivalent to separately estimating two vectors of OLS coefficients: $\hat{\theta}_1$ is estimated using only data from the treatment group, and $\hat{\theta}_0$ is estimated using only data from the control group. The estimator $\hat{\tau}$ can be recovered from these two regressions by first imputing the unobserved potential outcomes,

$$\hat{y}_{ti} = \begin{cases} y_{ti} & Z_i = t \\ \hat{\theta}_t x_i & Z_i \neq t \end{cases}$$

and then computing $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{ti} - \hat{y}_{0i})$. Since fitted values are not affected by centering, covariates can live on their original scale in these auxiliary regressions. In that sense, it is more natural to think of Lin’s “interactions” estimator as an imputation estimator. This viewpoint is also discussed by Ding and Li (2018) and Chapter 7 of the textbook by Imbens and Rubin (2015).

In econometrics, this double‐regression procedure is known as the Oaxaca–Blinder\(^3\) method (Blinder 1973; Oaxaca 1973; Kline 2011).

This perspective suggests a natural way of using an arbitrary regression model to estimate the sample average treatment effect: simply replace $\hat{\theta}_1 x_i$ in Equation (3) with $\hat{\mu}_1(x_i)$, where $\hat{\mu}_1$ is estimated (using any method) on the subset of observations

\(^1\)An R package exists, but implementing the method in STATA (for example) would still be challenging.
\(^2\)This tacitly assumes that there is no interference between experimental units, for example, the treatment assignment of unit $i$ does not affect the outcome of unit $j$ if $i \neq j$.
\(^3\)Although the “interactions” estimator is algebraically equivalent to the Oaxaca-Blinder estimator, Lin (2013) derived its asymptotic properties in a randomization-based framework—quite distinct from the inferential framework adopted in the econometrics literature.
with \( Z_i = t, t \in \{0,1\} \). This simple idea has been proposed (in one form or another) in many different research communities, but there does not seem to be a common name for it. We will call it the generalized Oaxaca-Blinder method—see Algorithm 1.

**Algorithm 1** The generalized Oaxaca-Blinder method

1. **Input.** Data \( \{(x_i, Y_i, Z_i)\}_{i=1}^{n} \).
2. Using data from treatment group, fit a regression model \( \hat{\mu}_1 \) that predicts \( y_{ti} \) using \( x_i \).
3. Use the model \( \hat{\mu}_1 \) to “fill in” the unobserved values of \( y_{ti} \) using (4). 
   \[
   \hat{y}_{ti} = \begin{cases} 
   y_{ti} & Z_i = 1 \\
   \hat{\mu}_1(x_i) & Z_i = 0 
   \end{cases} \tag{4}
   \]
4. Using data from control group, fit a regression model \( \hat{\mu}_0 \) that predicts \( y_{0i} \) using \( x_i \).
5. Use the model \( \hat{\mu}_0 \) to “fill in” the unobserved values of \( y_{0i} \).
   \[
   \hat{y}_{0i} = \begin{cases} 
   \hat{\mu}_0(x_i) & Z_i = 1 \\
   y_{0i} & Z_i = 0 
   \end{cases} \tag{5}
   \]
6. Return \( \hat{\tau} := \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{ti} - \hat{y}_{0i}) \).

Some historical remarks may be useful to put this algorithm in context. To our knowledge, the idea of using regression models to impute counterfactuals first appears in Shen (1940) and Peters (1941) in the educational literature, and it has subsequently been rediscovered in applied statistics (Belson 1956; Hansen and Bowers 2009), survey sampling (Särndal and Wirth 1984; Firth and Bennett 1998), economics (Oaxaca 1973; Blinder 1973; Fairlie 1999; Bauer and Sinning 2008; Kline 2011) and many other fields. In causal inference, Hahn (1998) and Abadie and Imbens (2006) have interpreted propensity-score matching estimators as counterfactual imputation methods, and Ferreira (2015) connected regression imputation to the G-computation formula of Robins (1986). Using this connection, many existing methods for covariate adjustment in randomized experiments (Rosenblum and van der Laan 2011; Bartlett 2018; Negi and Wooldridge 2020) can be interpreted as generalized Oaxaca-Blinder estimators.

Despite this rich history, not much is known about the statistical properties of the generalized Oaxaca-Blinder method in the randomization inference framework. Indeed, none of the aforementioned articles takes a finite-population view of regression imputation. In the Neyman model, rigorous results have only been established in the case of linear models, with some limited progress in the case of logistic regression (Freedman 2008c; Hansen and Bowers 2009).

### 2.2. Prediction Unbiasedness

Although any regression model can, in principle, be plugged into Algorithm 1, not every regression model will result in an estimate that is robust to misspecification. In the supplementary material, we give an example where a regression model that is substantially better than OLS in terms of predictive accuracy nevertheless yields a worse ATE estimate than Lin’s “interactions” estimator. The key property required to avoid such pathologies and ensure robustness to misspecification is prediction unbiasedness.

**Definition 1.** (Prediction unbiasedness) For \( t \in \{0,1\} \), we say that the regression model \( \hat{\mu}_t \) is prediction unbiased if (6) holds with probability one.

\[
\frac{1}{n_t} \sum_{i=1}^{n_t} \hat{\mu}_t(x_i) = \frac{1}{n_t} \sum_{i=1}^{n_t} y_{ti} \tag{6}
\]

In words, a regression model is prediction unbiased if the average prediction on the training data always exactly matches the average outcome in the training data. In survey sampling, this condition is called “calibration.”

Many widely used regression models are automatically prediction unbiased.1 For example, the first-order conditions of a canonical-link GLM imply Equation (6), so linear regression, logistic regression, and Poisson regression are prediction unbiased. Given an arbitrary prediction model \( \hat{\mu}_1 \), it is always possible to construct a related model \( \hat{\mu}_{1\text{db}} \) which is prediction unbiased by simply subtracting off the estimated bias as in Equation (7).

\[
\hat{\mu}_{1\text{db}}(x) = \hat{\mu}_1(x) - \frac{1}{n_1} \sum_{Z_i=1} \{\hat{\mu}_1(x) - y_{ti}\} \tag{7}
\]

Another possibility is to use the fitted values \( \hat{\mu}_1(x_i) \) as a covariate in an OLS regression as in Equation (8). The first-order condition characterizing \( \beta_0 \) in the least-square problem guarantees that \( \hat{\mu}_{1\text{sk2}} \) will also be prediction unbiased.

\[
\hat{\mu}_{1\text{sk2}}(x) = \hat{\beta}_0 + \hat{\beta}_1 \hat{\mu}_1(x),
\]

\[
(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(\beta_0, \beta_1)} \sum_{Z_i=1} (y_{ti} - \{\beta_0 + \beta_1 \hat{\mu}_1(x_i)\})^2 \tag{8}
\]

These are not the only ways of “debiasing” an initial prediction model \( \hat{\mu}_1 \). Another method which is more suitable for binary outcomes is discussed in Section 4.1. These examples are only used to illustrate that prediction unbiasedness does not seriously restrict the class of permissible nonlinear models.

When the debiasing trick (7) is used, the generalized Oaxaca-Blinder estimator is algebraically equivalent to the augmented inverse-propensity weighted treatment effect estimator with known (and constant) treatment propensity (Robins, Rotnitzky, and Zhao 1994b). In the survey sampling literature, that estimator is known as the generalized difference estimator (Casel, Särndal, and Wretman 1976; Breidt and Opsomer 2017). The idea of using fitted values in an OLS regression has also appeared before in the survey sampling community, under the name “model calibration estimator” (Wu and Sitter 2001), and a similar idea appears in Chakrabortty and Cai (2018). Although those connections are mathematically fruitful, we believe that the formulation we present ("estimate \( \tau \) by filling in missing values with unbiased prediction models") is much more intuitive in the context of Neyman’s finite-population model.

---

1 Such estimators are called “internally bias calibrated” by Firth and Bennett (1998).
2.3. Statistical Inference

Under some additional constraints on the regression models $\hat{\mu}_1$ and $\hat{\mu}_0$ (to be discussed in Section 3), the confidence interval (9) has large-sample validity.

$$\hat{r} \pm z_{1-\alpha/2} \sqrt{\frac{\text{MSE}(1)}{n_1} + \frac{\text{MSE}(0)}{n_0}}$$

(9)

In the above display, $\text{MSE}(t) = \frac{1}{n_t} \sum_{i=1}^{n_t} (y_{it} - \hat{\mu}_t(x_i))^2$ is an estimate of the mean-squared error of the prediction model $\hat{\mu}_t$. The form of the confidence interval (9) has some intuitive appeal: when more accurate models are used to “fill in” the missing values, the resulting estimator $\hat{r}$ is more precise.

When the regression models are constant (i.e., $\hat{\mu}_t(x) \equiv \frac{1}{n_t} \sum Z_{i=1} y_{it}$), we recover the confidence interval suggested by Neyman (1923) for the difference-of-means estimator. The robust standard error confidence interval based on Lin’s “interactions” estimator is—in large samples—statistically equivalent to the interval (9) when both regression models are linear models. Therefore, when nonlinear regression models fit the data better than linear models, the interval (9) is shorter than the “robust” standard error interval from OLS adjustment.

2.4. An Illustration

Before going into theoretical details, we briefly illustrate the computational and numerical properties of the generalized Oaxaca-Blinder estimator with a simple example. The Fatalities dataset in R was introduced by Ruhm (1995), and contains the number of traffic fatalities in each continental U.S. state between 1982 and 1988 along with a few covariates. To study the effect of a (fictional) randomized intervention designed to reduce traffic fatalities, Lin’s “interactions” estimator is a natural baseline. However, statistical intuition suggests that Poisson models might fit better, since the outcome variable counts the occurrences of a rare event.

Computing the generalized Oaxaca–Blinder based on Poisson regression and its associated confidence interval using Algorithm 1 and formula (9) is not too difficult, but there are two observations that can make it even simpler. First, for prediction unbiased regression models (like Poisson regression), the right-hand side of the identity (10) is often easier to work with.

$$\hat{r} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{1i} - \hat{y}_{0i}) = \frac{1}{n} \sum_{i=1}^{n} [\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)]$$

(10)

Second, the interval (9) can be computed by simply adding $\hat{r}$ to the endpoints of the confidence interval from a two-sample $t$-test\(^6\) on the residuals of $\hat{\mu}_0$ and $\hat{\mu}_1$. Using these two computational shortcuts, the estimator and its confidence interval can be computed in only four lines of R code.

For comparison, computing Lin’s “interactions” estimator in R requires roughly the same amount of code, at least when the features include at least one factor variable. In fact, it may be easier for some users to use this method to compute $\hat{r}$ even when $\hat{\mu}_1$ and $\hat{\mu}_0$ are linear models, because centering factors in STATA/SAS/Excel is nonstandard.

We ran the above code 50,000 times on the Fatalities dataset, rerandomizing the treatment assignments ($Z_1, Z_{365}$) in each replication. Each time, we also computed the robust standard error confidence interval based on Lin’s “interactions” estimator. In our simulations, both potential outcomes $y_{it}$ and $y_{i0}$ were the same for each experimental unit, so the true average treatment effect is zero.\(^7\) Figure 1 plots the randomization distributions of these two estimators. Two features are immediately clear: (i) both estimators have an approximately normal randomization distribution; (ii) the Poisson regression generalized Oaxaca-Blinder estimator is much more efficient than the “interactions” estimator. This efficiency gain is also reflected in the width of the associated confidence intervals: 95% confidence intervals based on Poisson imputation were about 45% shorter (on average) than the robust standard error confidence intervals\(^8\) based on Lin’s “interactions” estimator. Both confidence intervals had approximately nominal coverage.

3. Theoretical Results

In this section, we state our main theoretical results concerning the consistency and asymptotic normality of generalized Oaxaca-Blinder estimators. The assumptions in this section are deliberately high-level, since the results are intended to cover a wide variety of examples. More low-level assumptions are used to specialize these results to specific regression methods in Section 4.

Like the results of prior work on covariate adjustment with linear models (Bloniarz et al. 2016; Fogarty 2018; Freedman 2008a,b; Li and Ding 2020; Liu and Yang 2020), our theoretical guarantees are asymptotic. In Neyman’s finite-population model, this means triangular-array asymptotics with respect to a sequence of finite populations $\Pi_n = \{(y_{1i, n}, y_{0i, n}, x_{in})\}_{i=1}^{n}$ of increasing size, each with its own completely randomized experiment $(Z_{1i, n}, \ldots, Z_{ni, n}) \sim F_{ni, n}$ and treatment effect $\tau_n = \frac{1}{n} \sum_{i=1}^{n} (y_{1i, n} - y_{0i, n})$. We focus on the low-dimensional regime where $d$ stays fixed as $n$ grows. Although we do not assume that

\footnotesize
\begin{enumerate}
\item The right-hand side of Equation (10) is reminiscent of a “marginal effects” calculation, and of the “G-formula.” We have not pursued that connection because the interpretation does not make sense in the randomization model.
\item This changes the normal quantile $z_{1-\alpha/2}$ to a $t$-quantile, which we recommend.
\item The reason for looking at a simulation with no individual treatment effect rather than a simulation with merely no average treatment effect is that confidence intervals are conservative under the latter, and simulations are therefore less informative. See the comment at the beginning of Section 7.2 in Lin (2013)
\item We used the “HC3” version of the robust standard errors. The “HC0” standard errors studied in Lin (2013) did not have good coverage properties in this example.
\end{enumerate}
these experiments are related in any way, we will assume in what follows that the fraction of treated units \( p_n = n_{1,n}/n \) satisfies \( 0 < p_{\text{min}} \leq p_n \leq p_{\text{max}} < 1 \) for some bounds \( p_{\text{min}}, p_{\text{max}} \) that do not vary with \( n \).

Some remarks on notation: for simplicity, we will often drop the \( n \)-subscript on various quantities, for example, we will write \( x_i \) in place of \( x_{i,n} \). For two symmetric matrices \( A, B \in \mathbb{R}^{d \times d} \), we will write \( A \succeq B \) if \( A - B \) is positive semidefinite. For any functions \( f, g : \mathbb{R}^d \to \mathbb{R} \), we define \( ||f - g||_n = (\frac{1}{n} \sum_{i=1}^n (f(x_{i,n}) - g(x_{i,n}))^2)^{1/2} \). Throughout, we use \( c, c', C, C', \ldots \) to refer to finite constants that do not change with \( n \) (but may depend on the population sequence \( \{\Pi_n\}_{n \geq 1} \)). Their precise value may change from line to line.

### 3.1. Consistency

Under very weak conditions, generalized Oaxaca-Blinder estimators based on prediction-unbiased regression models are consistent in the randomization model. The regression models \( \hat{\mu}_0 \) and \( \hat{\mu}_1 \) do not need to be correctly specified for this to hold. In biostatistics and survey sampling, models which are used to construct an estimator but which are not believed to be correct are called “working models.” We will adopt this terminology as well.

To build some intuition for why correct specification is not required for consistency, consider the Poisson regression working model presented in Section 2.4. The regression models are of the form \( \hat{\mu}_t(x) = \exp(\hat{\theta}_t^\top x) \), where \( \hat{\theta}_t \) solves Equation (11).

\[
\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \sum_{Z_i=t} \left[ -y_i x_i^\top \theta + \exp(\theta^\top x_i) \right] \quad (11)
\]

Since the subset of observations with \( Z_i = t \) is a random sample of all the experimental units, we would expect that \( \hat{\theta}_t \) is close to the solution of the population version of the problem (11), where the sum is taken over all \( i \) instead of only those with \( Z_i = t \). Let \( \theta_t^* \) be the solution of the “population” problem. When the covariates \( x_i \) include an intercept (which we will assume throughout this article), the first-order condition characterizing \( \theta_t^* \) implies Equation (12).

\[
\frac{1}{n} \sum_{i=1}^n \exp(\theta_t^*^\top x_i) = \frac{1}{n} \sum_{i=1}^n y_i \quad (12)
\]

Therefore, we could (heuristically) argue

\[
\frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n \left( \sum_{Z_i=t} y_i + \sum_{Z_i \neq t} \exp(\hat{\theta}_t^\top x_i) \right) \\
= \frac{1}{n} \sum_{i=1}^n \left( \sum_{Z_i=t} \exp(\hat{\theta}_t^\top x_i) + \sum_{Z_i \neq t} \exp(\hat{\theta}_t^\top x_i) \right) \\
\approx \frac{1}{n} \sum_{i=1}^n \exp(\theta_t^*^\top x_i) \\
= \frac{1}{n} \sum_{i=1}^n y_i \quad \text{(Prediction unbiased)}
\]

Since this argument works for both \( t = 0 \) and \( t = 1 \), we have \( \hat{\tau} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \hat{\mu}_0) \approx \frac{1}{n} \sum_{i=1}^n (y_i - \mu_0) = \tau \).

This simple argument is well known in the survey sampling community (Firth and Bennett 1998; Särndal and Wright 1984), and it is the main idea behind Freedman’s consistency result for logistic regression (Freedman 2008b). It is almost completely rigorous. The only step that needs to be justified is the claim that \( \exp(\hat{\theta}_t^\top x_i) \approx \exp(\theta_t^*^\top x_i) \) (at least on average). A sufficient\(^9\) condition to make this argument rigorous is stability.

\(^9\) This condition is not necessary for consistency, but it plays a key role in our later results on asymptotic normality.
Definition 2. (Stability) We say that a sequence of random functions \( \{ \hat{\mu}_n \}_{n \geq 1} \) is stable if (13) holds for some deterministic sequence of functions \( \{ \mu^*_n \}_{n \geq 1} \).

\[
||\hat{\mu}_n - \mu^*_n||_n^2 := \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_n(x_{i,n}) - \mu^*_n(x_{i,n}))^2 \rightarrow_p 0 \tag{13}
\]

Remark 1. In words, a random function \( \hat{\mu}_n \) is stable if its predictions are roughly the same as those of some fixed regression model. If the sequence of finite populations \( (\Pi_{n})_{n \geq 1} \) is obtained by iid sampling from an infinite superpopulation (i.e., \( \Pi = \{ (Y_{1i}, Y_{0i}, X_i) \}_{i=1}^{\infty} \) with \( (Y_{1i}, Y_{0i}, X_i) \sim \mathcal{D} \)), then Equation (13) will hold with probability one as long as the regression model converges to some limit as \( n \to \infty \). The limit does not need to be “correct” in any sense. In that case, our definition is similar to (but weaker than) the notion of predictor stability proposed by Bühlmann and Yu (2002).

The deterministic sequence in the definition of stability is not uniquely determined, but there is usually a natural choice. For example, if \( \hat{\mu}_n = \mu_{\beta_0} \) is a parametric regression model estimated via maximum likelihood, then \( \mu^*_n = \mu^{\text{OLS}}_{\beta_0} \) is the clear candidate. For this reason, we will typically call \( \mu^*_n \) “the” population regression function, even without specifying exactly which choice of \( \mu^*_n \) we are making. Perhaps surprisingly, the definition of stability does not assume that the deterministic sequence \( \{ \mu^*_n \}_{n \geq 1} \) satisfies a property like Equation (12). It turns out that if \( \{ \hat{\mu}_n \}_{n \geq 1} \) is both prediction unbiased and stable, then we may always choose the sequence \( \{ \mu^*_n \}_{n \geq 1} \) to satisfy

\[
\frac{1}{n} \sum_{i=1}^{n} \mu^*_n(x_{i,n}) = \frac{1}{n} \sum_{i=1}^{n} y_{1i} \quad \text{Therefore, prediction unbiasedness and stability are sufficient to carry through the heuristic argument from above—Theorem 2 gives a formal statement.}
\]

Theorem 2. (Consistency) Let \( \{ \hat{\mu}_{1,n} \}_{n \geq 1} \) and \( \{ \hat{\mu}_{0,n} \}_{n \geq 1} \) be two stable sequences of prediction-unbiased models, and let \( \{ \hat{\tau}_n \}_{n \geq 1} \) be the resulting sequence of generalized Oaxaca-Blinder estimators. For \( t \in [0,1] \), let MSE\(_n\)(t) := \( \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_{tin}(x_{i,n}) - y_{1in})^2 \) be the mean-squared error of the population model \( \mu_{tin} \). If MSE\(_n\)(1) = o(n) and MSE\(_n\)(0) = o(n), then \( \hat{\tau}_n \to \tau \) a.s.

Note that Theorem 2 does not depend on the specific choice of the sequences \( \{ \mu^*_{1,n} \}_{n \geq 1} \) and \( \{ \mu^*_{0,n} \}_{n \geq 1} \). As long as any non-random sequence \( \{ \mu^*_{tin} \}_{n \geq 1} \) satisfying \( ||\hat{\mu}_{tin} - \mu^*_{tin}||_n \to_p 0 \) has MSE\(_n\)(t) = o(n), then all such sequences will have that property. In most cases, the mean-squared error of even a grossly misspecified model is not diverging at all, so we would have MSE\(_n\)(1) = O(1), MSE\(_n\)(0) = O(1). We have used the weaker assumption in Theorem 2 only for the sake of generality.

Proving that a sequence of random functions \( \{ \hat{\mu}_n \}_{n \geq 1} \) is stable is typically an exercise in translating some standard arguments from the theory of M-estimation into the language of finite populations. This may or may not be simple, depending on the regression function. In Section 4, we give a few examples of widely-used regression methods where this can be done, including logistic regression, Poisson regression, and OLS regression with a transformed outcome variable. In Section 5, we outline a general strategy that works for a large class of smooth parametric models. The argument is especially simple in the case of linear regression, so we present it as an example.

Example 1. (OLS is stable) Let \( \hat{\mu}_{1,n}(x) = \hat{\beta}_1^\top x, \) where \( \hat{\beta}_1 = \text{argmin} \sum_{i=1}^{n} (y_{1in} - x_{i,n}^\top \hat{\beta}_1)^2 \).

Assume that \( \frac{1}{n} \sum_{i=1}^{n} ||\hat{\beta}_1||^4 = o(n) \), \( \frac{1}{n} \sum_{i=1}^{n} y_{1i}^2 = o(n) \), and \( \frac{1}{n} \sum_{i=1}^{n} (y_{1i}, x_{i,n}(y_{1i}, x_{i,n})^\top \) converges to an invertible matrix. Then \( \{ \hat{\mu}_{1,n} \}_{n \geq 1} \) is a stable sequence.

Proof. We can write \( \hat{\beta}_{1,n} = (\frac{1}{n} \sum_{i=1}^{n} x_{i,n} x_{i,n}^\top)^{-1}(\frac{1}{n} \sum_{i=1}^{n} x_{i,n} y_{1i}) \).

Let \( \Sigma_{xx} \) be the limit of \( \frac{1}{n} \sum_{i=1}^{n} x_{i,n} x_{i,n}^\top \) and \( \Sigma_{sy} \) be the limit of \( \frac{1}{n} \sum_{i=1}^{n} x_{i,n} y_{1i} \). By the completely randomized law of large numbers and the continuous mapping theorem, \( \hat{\beta}_{1,n} \to_p \beta_1 = \Sigma_{xx}^{-1} \Sigma_{sy} \). If we set \( \mu^*_{1,n}(x) = \beta_1^\top x \), then we may write

\[
||\hat{\mu}_{1,n} - \mu^*_{1,n}||_n^2 \leq \frac{1}{n} \sum_{i=1}^{n} ||(\hat{\beta}_1 - \beta_1^\top x_i)||^2 \leq \text{Tr}(\Sigma_{xx}) ||\hat{\beta}_1 - \beta_1^\top||_F^2 \to_p 0
\]

Thus, \( \{ \hat{\mu}_{1,n} \}_{n \geq 1} \) is stable. \( \square \)

3.2. Asymptotic Normality

In order prove the asymptotic normality of generalized Oaxaca-Blinder estimators, we require that the working models \( \hat{\mu}_1 \) and \( \hat{\mu}_0 \) be “simple”.

Definition 3. (Simple prediction models) We say that the sequence \( \mu_{tin} \) is typically simple or has bounded entropy if, for some sequence of function classes \( F_n \), the following hold:

1. “Typical” \( \mathbb{P}(\hat{\mu}_n \in F_n) \to 1 \).
2. “Simple.” The metric entropy \(^{10}\) of \( F_n \) satisfies \( \log N(F_n, || \cdot ||_{n,s}) \leq C/s^k \) for some \( k < 2 \).

Remark 2. In the “superpopulation” model of causal inference, entropy bounds are a standard ingredient in asymptotic normality proofs. See Rothe (2018), Kennedy (2016), and the discussion in Chernozhukov et al. (2018). For stable sequences \( \{ \hat{\mu}_{tin} \}_{n \geq 1} \), these bounds will be satisfied as long as \( \hat{\mu}_{tin} \) is “simple” enough to not overfit the data. Examples include smooth parametric models of fixed dimension, classifiers in a VC class, regression models with sufficiently many bounded derivatives, and linear combinations of such functions. Kosorok (2008) and van der Vaart and Wellner (1996) gave a wealth of additional examples. To our knowledge, entropy bounds have not previously been used in randomization inference. While such bounds immediately enable the application of empirical process maximal inequalities in the case of iid sampling, the corresponding finite-population inequalities were not available. Indeed, the exponential tail bounds needed to derive such results were introduced to the this literature only very recently (Bloniarz et al. 2016; Wu and Ding 2020).

\(^{10}\)The s-covering number of a metric space \( (F_n, || \cdot ||_n) \), denoted by \( N(F_n, || \cdot ||_{n,s}) \), is the size of the smallest collection \( \{ f_1, \ldots, f_n \} \subseteq F_n \) such that every \( f \in F_n \) is within distance \( s \) of one of the \( f_i \)'s. The metric entropy is the logarithm of the covering number.
Although Definition 3 is a somewhat technical definition, it is fairly easy to check. For illustration, we show how it can be established in the case of OLS regression.

Example 2. (OLS has typically simple realizations). Let \( \hat{\beta}_{1,n}(x) = \beta_1^* x \), and assume that the conditions in Example 1 are satisfied. Then \( \{\hat{\beta}_{1,n}\}_{n \geq 1} \) has typically simple realizations.

**Proof.** Since \( \hat{\beta}_{1,n} \to_p \beta_1^* \), the function \( \hat{\beta}_{1,n} \) typically takes values in the “simple” set \( F_n = \{\mu_\beta(x) : ||\beta - \beta_1^*|| \leq 1\} \). To see that this set is simple, we use the fact for all \( \beta, \gamma \), \( ||\mu_\beta - \mu_\gamma|| \leq C||\beta - \gamma|| \) for some \( C < \infty \).

\[
||\mu_\beta - \mu_\gamma||_n^2 = \frac{1}{n} \sum_{i=1}^{n} (\beta - \gamma)^\top x_i^2 \\
\leq \text{Tr}(\Sigma_n)||\beta - \gamma||^2 \leq C^2||\beta - \gamma||^2.
\]

The last inequality is valid with \( C^2 = 2\text{Tr}(\Sigma) \) for large enough \( n \), since \( \Sigma_n \to \Sigma \). As a consequence, the \( s \)-covering number of \( F_n \) can be bounded using the \((s/C)\)-covering number of the Euclidean ball \( B_1(\beta_1^*) \) \( : \{\beta \in \mathbb{R}^d : ||\beta - \beta_1^*|| \leq 1\} \). A simple volume argument shows that the \( s \)-covering number of \( B_1(\beta_1^*) \) is less than or equal to \((1 + 2/s)^d\), so the metric entropy of \( F_n \) is bounded by \( (1 + 2C/s)^d \leq 2Cd/s \).

The argument above works whenever \( \hat{\theta} \to \theta^* \) for some limit \( \theta^* \) and \( \theta \to \mu_\theta \) is smooth near \( \theta^* \). More general nonparametric function classes can also be shown to have the bounded entropy property, using combinatorial arguments.

If simple, stable, prediction-unbiased models are used in the imputation step of the generalized Oaxaca-Blinder method, then the imputed means \( \frac{1}{n} \sum_{i=1}^{n} \hat{y}_{ii} \) and \( \frac{1}{n} \sum_{i=1}^{n} \hat{y}_{0i} \) both have asymptotically linear expansions. This is stated formally in Theorem 3.

**Theorem 3. (Asymptotically linear expansion)** Let \( \{\hat{\mu}_{1,n}\}_{n \geq 1} \) and \( \{\hat{\mu}_{0,n}\}_{n \geq 1} \) satisfy the assumptions of Theorem 2, and further suppose that these models are simple. Then we have the following asymptotically linear expansions

\[
\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{ii} - y_{ii}) = \frac{1}{n_1} \sum_{Z_{i}=1} e_{1i} + o_p(n^{-1/2}), \quad (14) \\
\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{0i} - y_{0i}) = \frac{1}{n_0} \sum_{Z_{i}=0} e_{0i} + o_p(n^{-1/2}), \quad (15)
\]

where \( e_{1i} := y_{ii} - \mu_{1,n}(x_i) \) and \( e_{0i} := y_{0i} - \mu_{0,n}(x_i) \). Moreover, \( \mu_{1,n}^* \) and \( \mu_{0,n}^* \) may be chosen so that \( \frac{1}{n} \sum_{i=1}^{n} e_{1i}^2 = \frac{1}{n} \sum_{i=1}^{n} e_{0i}^2 = 0 \).

In words, Theorem 3 says that \( \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{ii} - y_{ii}) \) is essentially just a sample average of some mean-zero constants. Thus, under some Lindeberg-type conditions, the completely randomized central limit theorem (Li and Ding 2017) can be used to prove the joint asymptotic normality of \( \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{ii} - y_{ii}) \) and \( \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{0i} - y_{0i}) \), which implies the asymptotic normality of \( (\hat{\tau}_n - \tau_0) \). Corollary 1 gives some sufficient conditions.

**Corollary 1. (Asymptotic normality)** Assume the conclusion of Theorem 3. Define \( \sigma_n^2 := \frac{1}{n_1} \text{MSE}_n(1) + \frac{1}{n_0} \text{MSE}_n(0) - \frac{1}{n} \sum_{i=1}^{n} (e_{1i}^* - e_{0i}^*)^2 \). Assume that \( \sigma_n^2 \) is bounded away from zero and \( \max_i (e_{1i}^*)^2 = o(n) \) for \( t \in \{0, 1\} \). Then we have:

\[
\frac{\sqrt{\hat{\tau}_n - \tau_0}}{\sigma_n} \to_d N(0, 1). \quad (16)
\]

Corollary 1 holds for any choice of the sequences \( \{\mu_{1,n}^*\}_{n \geq 1} \) and \( \{\mu_{0,n}^*\}_{n \geq 1} \) that satisfy the stated conditions. The requirement that \( \sigma_n^2 \) is bounded away from zero is only used to rule out the degenerate situation where \( \sqrt{\hat{\tau}_n - \tau_0} = o_p(1) \). In that case, \( \hat{\tau}_n \) is still a very good estimate of \( \tau_0 \), but asymptotic normality may not hold and confidence intervals may not have valid coverage. The reason is that if the linear terms in Equations (14) and (15) are of comparable magnitude to the \( o_p(n^{-1/2}) \) remainder terms, then the remainder terms may contribute to the asymptotic distribution. Even if they do not contribute, confidence interval coverage may be compromised by the difficulty of accurately estimating \( \sigma_n \) when \( \sigma_n \to 0 \). Although this condition is not used in the works by Lin (2013) and Bloniarz et al. (2016), that seems to be an oversight. In practice, users should check the \( R^2 \) from their regression models; if they are both very close to one, there may be reason for concern.

Unlike Lin’s result (Theorem 1), Corollary 1 does not contain any “noninferiority” claim; that is, there is no guarantee that a generalized Oaxaca-Blinder estimator is never worse than Neyman’s unadjusted difference-of-means estimator in terms of its sampling variability. That is the price of generality. Since the class of regression methods to which Corollary 1 applies is so broad, it inevitably contains some bad apples. See Section 4 in Cohen and Fogarty (2020) for an example. In practice, however, it is quite difficult to construct an explicit example where a regression method that is actually used has worse performance than the unadjusted estimator.

**3.3. Confidence Intervals**

In order to use Corollary 1 to construct confidence intervals for \( \tau_0 \), it is necessary to construct an estimate of the standard error \( \sigma_n \). Although \( \sigma_n \) is not (in general) identifiable from observed data, there is an identifiable upper bound which follows from the calculations by Neyman (1923)

\[
\text{bias}^2 \leq \frac{1}{n_1} \text{MSE}_n(1) + \frac{1}{n_0} \text{MSE}_n(0). \quad (17)
\]

One way to estimate the upper bound in Equation (17) is to first use in-sample residual variance (18) as an estimate of \( \text{MSE}_n(t) \), and then plug the MSE estimates back into Equation (17). Then, confidence intervals may be constructed using the estimated upper bound:

\[
\widetilde{\text{MSE}}_n(t) := \frac{1}{n_t - 1} \sum_{Z_{i}=t} (\hat{y}_{ii} - \hat{\mu}_{t,n}(x_i))^2. \quad (18)
\]
Theorem 4 says that this will work, as long as the residuals from the “population” model $\mu^*_1$ and $\mu^*_0$ are not too heavy-tailed.

**Theorem 4.** (Confidence intervals)
Assume the conditions of Theorem 3 and Corollary 1. If MSE$_n(t)$ stays bounded and $\frac{1}{n} \sum_{i=1}^{n} (\epsilon_i^*)^4 = o(n)$ for $t \in \{0, 1\}$, then we may construct an asymptotically valid confidence interval for $\tau_n$

$$\lim_{n \to \infty} \Pr \left[ \hat{\tau}_n \in \left[ \hat{\tau}_n \pm z_{1-\alpha/2} \sqrt{\frac{\text{MSE}_n(1)}{n_1} + \frac{\text{MSE}_n(0)}{n_0}} \right] \right] \geq 1 - \alpha. \quad (19)$$

When the treatment has no effect (i.e., $y_{1i} = y_{0i}$ for all $i$) and the same method is used to estimate both $\mu_1$ and $\mu_0$ (e.g., both logistic regression with the same covariates), then the asymptotic coverage of the confidence interval is exactly $1 - \alpha$.

Our proposed confidence intervals have a certain “non-inferiority” property when the working models $\mu_0$ and $\mu_1$ are estimated by least-squares: they are never wider than Neyman’s unadjusted confidence intervals. This applies not only to OLS adjustment, but also to the “OLS2” estimator (8) based on an arbitrary initial nonlinear estimator. It also applies to nonparametric least-squares estimators (see Section 4.4 for an example). The explanation is that Neyman’s confidence intervals are exactly of the form Equation (9), with $\hat{\sigma}_i^2 = \frac{1}{n_{i-1}} \sum_{i=t} Z_{i=t} (y_{ii} - \hat{y}_{ii})^2$ in place of $\text{MSE}(t)$. If $\hat{\mu}_1$ is estimated by least-squares, then MSE$(t) \leq \hat{\sigma}_i^2$ holds deterministically as long as $\hat{\mu}_i$ belongs to a function class containing all constant functions. However, as discussed above, this “non-inferiority” property of CI widths does not imply a non-inferiority property in terms of actual sampling variability. See Li and Ding (2020) for a more detailed discussion on the relationship between these two forms of non-inferiority.

### 3.4. Extensions

The results presented above have a host of other consequences, some of which we will briefly allude to here.

First, the asymptotically linear expansions of Theorem 3 hold under any experimental design that is contiguous to the completely randomized design, including the Mahalanobis randomization scheme studied in Li and Ding (2020). Therefore, using the technical tools in that work, one can develop asymptotic theory for generalized Oaxaca-Blinder estimators in randomized designs as well.

Second, recent work by Zhao and Ding (2020) showed that using a studentized version of Lin’s “interactions” estimator as the test statistic in Fisher’s randomization test (FRT) gives p-values that are exact for Fisher’s “sharp” null hypothesis (c.f. Imbens and Rubin 2015) and asymptotically valid for Neyman’s “weak” null hypothesis $\tau = 0$. Since the work of Chung and Romano (2013) obtained a similar conclusion for general asymptotically linear test statistics, it seems likely that the result could be extended to generalized Oaxaca–Blinder estimators.

We leave the technical details of these extensions for future work.

### 4. Examples

In this section, we give some examples of specific regression methods that satisfy the assumptions of stability, prediction unbiasedness, and typically simple realizations. We have chosen examples that cover a variety of situations in which Lin’s “interactions” estimator might be deficient: (i) binary outcomes; (ii) count outcomes; (iii) skewed outcomes; and (iv) highly nonlinear relationships.

#### 4.1. Logistic Regression

With binary outcomes, OLS covariate adjustment is intuitively “wrong.” For example, if the perspective on Lin’s “interaction” estimator presented in Section 2 is taken up, imputation with linear models is unsatisfactory because values less than 0 or larger than 1 may be imputed. This deficiency has led many authors to consider procedures similar to the logistic-regression-based Oaxaca-Blinder estimator (Firth and Bennett 1998; Freedman 2008c; Hansen and Bowers 2009; Ding and Li 2018), but none have given a truly satisfying proof that covariate adjustment with logistic models is valid in the randomization model.

The closest result we have seen is the work of Hansen and Bowers (2009), which established the asymptotic normality of $\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{1i} - y_{1i})$ when $\hat{y}_{1i}$ is imputed using logistic regression. However, the authors simply assumed stability instead of proving it from low-level assumptions on the finite population. This appears to be a common theme in the finite-population analysis of generalized linear models (Binder 1983; Roberts, Rao, and Kumar 1987; Fuller 2009). Theorem 5 goes beyond that and establishes the asymptotically linear expansion for $\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{1i} - y_{1i})$ under only primitive assumptions on the population. Asymptotic normality follows under the additional requirements of Corollary 1.

**Theorem 5.** (Logistic regression)

Suppose that the potential outcomes $y_{1i}$ are binary and $\hat{\mu}_1$ is estimated using logistic regression. Assume that, for all large $n$, there exists a vector $\theta_{1,n}$ solving the “population” logistic regression problem (20).

$$\theta_{1,n} = \arg\min_{\theta} \left\{ \mathcal{L}_n(\theta) : -\frac{1}{n} \sum_{i=1}^{n} -y_{1i} x_i^T \theta + \log(1 + e^{\theta^T x_i}) \right\}.$$

(20)

Further suppose that $\frac{1}{n} \sum_{i=1}^{n} ||x_i||^4$ is uniformly bounded and $\nabla^2 \mathcal{L}_n(\theta_{1,n}) \succeq c I_{d \times d}$ for some $c > 0$. Then the sequence $(\hat{\mu}_{1,n})_{n \geq 1}$ satisfies the conditions of Theorem 3.

At least qualitatively speaking, the assumptions of Theorem 5 cannot be improved (although the uniform bounds could, with additional effort, be relaxed to depend on $n$). It is widely known that the logistic MLE fails to exist when the design matrix $X$ is rank-deficient or the 1s and 0s among the outcomes can be perfectly separated by a hyperplane (Albert and Anderson 1984).

If either of these properties hold in the population, then they necessarily hold in a subsample. The existence of the population MLEs is, therefore, necessary to guarantee the existence of the sample MLEs.
However, it is not sufficient. Consider for instance the following possibilities:

- **Near-perfect separation.**
  If only a single exceptional point prevents the $y_i$’s from being perfectly separated by a hyperplane, then $\theta_i^* = 0$ exists. However, in $100(1 - p_n)\%$ of the possible realizations of the treatment assignments $(Z_1, \ldots, Z_n)$, the exceptional point will not be observed. In those samples, $\hat{\theta}_1$ will fail to exist and $\hat{\gamma}$ will not be well-defined.

- **Near-perfect collinearity.**
  If two columns of the design matrix $X$ have perfect correlation except for a few exceptions, then in many realizations of the treatment assignments $(Z_1, \ldots, Z_n)$, the design matrix used to estimate $\hat{\mu}_1$ will be rank-deficient. Again, $\hat{\theta}_1$ will fail to exist. This might occur if one of the predictors is a very sparsely populated indicator variable.

The assumption that the Fisher information $\nabla^2 \mathcal{L}_n(\theta_{1,n}^*)$ is bounded away from zero is used to rule out such pathological populations. Some indications that these assumptions are violated are (i) many fitted values very close to 0.00 or 1.00, and (ii) “unnaturally large” standard errors on the regression coefficients.

As a final comment on logistic regression, we will point out that one obtains fitted values $\hat{y}_i = g^{-1}(\hat{\theta}^\top x_i)$, i.e., back-transform to obtain fitted values on the original scale. In general, $\hat{\mu}_1$ will not be prediction unbiased, but this can be corrected using the “debiased” prediction model $\hat{\mu}_1^{db}(x) = \hat{\mu}(x) - \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu}(x_i))$ (22) or the “OLS2” prediction model

$$
\hat{\mu}_1^{ols2}(x) = \hat{\beta}_0 + \hat{\beta}_1 \hat{\mu}(x),
$$

$$
(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(\beta_0, \beta_1)} \frac{1}{n} \sum_{i=1}^n (y_i - [\beta_0 + \beta_1 \hat{\mu}(x_i)])^2.
$$

Finally, either $\hat{\mu}_1^{db}$ or $\hat{\mu}_1^{ols2}$ can be used as the basis for a generalized Oaxaca-Blinder estimator. Theorem 7 says that, under assumptions similar to those used in the case of logistic and Poisson regression, either of these two prediction models will “work”. On the other hand, the supplementary materials give an example showing that skipping the calibration steps and naively plugging the back-transformed estimator $\hat{\mu}(x)$ directly into Algorithm 1 does not work.

**4.3. Transformed-Outcome Regression**

Our next example covers cases where a linear model is more appropriate after applying a transformation to the outcome variable. The main examples we have in mind are the log transformation for skewed outcomes and the logit transformation for bounded outcomes. More generally, the results will apply with any invertible outcome transformation $g(\cdot)$ whose inverse has a continuous, positive derivative.

One simple way of analyzing such data is transform the outcome variable and then apply Lin’s “interactions” estimator. Indeed, this was the strategy used in the original article by Oaxaca (1973), which modeled log wages. However, since treatment effects are generally more interpretable on the original scale (e.g., dollars instead of log dollars), this is not entirely satisfying.

To get around this problem, one can adopt the following procedure. First, regress $g(y_{1i})$ on $x_i$ to obtain a vector of regression coefficients $\hat{\beta}_1$. Then, define $\hat{\mu}_1(x) = g^{-1}(\hat{\theta}^\top x)$, that is, back-transform to obtain fitted values on the original scale. In general, $\hat{\mu}_1$ will not be prediction unbiased, but this can be corrected using either the “debiased” prediction model $\hat{\mu}_1^{db}(x) = \hat{\theta}(x) - \frac{1}{n} \sum_{i=1}^n (y_{1i} - \hat{\theta}(x_i))$ (22) or the “OLS2” prediction model

$$
\hat{\mu}_1^{ols2}(x) = \hat{\phi}_0 + \hat{\phi}_1 \hat{\mu}(x),
$$

$$
(\hat{\phi}_0, \hat{\phi}_1) = \arg\min_{(\phi_0, \phi_1)} \frac{1}{n} \sum_{i=1}^n (y_{1i} - [\phi_0 + \phi_1 \hat{\mu}(x_i)])^2.
$$

Finally, either $\hat{\mu}_1^{db}$ or $\hat{\mu}_1^{ols2}$ can be used as the basis for a generalized Oaxaca-Blinder estimator. Theorem 7 says that, under assumptions similar to those used in the case of logistic and Poisson regression, either of these two prediction models will “work”. On the other hand, the supplementary materials give an example showing that skipping the calibration steps and naively plugging the back-transformed estimator $\hat{\mu}(x)$ directly into Algorithm 1 does not work.
Although the “OLS2” model requires the mild additional assumption that at least one predictor is correlated with the transformed outcome, we nevertheless recommend it over the “debiased” model. The reason is that the extra degree of freedom \( \hat{\beta}_1 \) can make a big difference in terms of model fit. For example when the “true” data-generating process is \( y_{1i} = \exp(\theta_1^\top x_i) + \epsilon_i \) where \( \epsilon_i \sim \mathcal{N}(0, \sigma^2) \), then the optimal predictor is \( \beta_1 \exp(\theta_1^\top x_i) \) where \( \beta_1 = \exp(\sigma^2/2) \).

We illustrate the difference using an example. We used the dataset and experimental set-up from Example 2.4 to compare the two transformed-outcome methods. The outcome in this dataset (number of traffic fatalities) is highly skewed, and linear models fit much better after a log transformation of outcomes than the (in-sample) mean-squared error from isotonic regression.

The theoretical properties of isotonic regression in the iid setting are well-studied (see the review article, Guntuboyina and Sen 2018), but no previous works have studied this method from the perspective of randomization inference. Theorem 8 says that, under almost no assumptions, isotonic regression satisfies all of the conditions needed to use it in the generalized Oaxaca-Blinder procedure.

\[ \hat{\mu}_1 \in \arg\min_{\mu \in \mathcal{M}} \frac{1}{n_1} \sum_{i=1}^{n_1} (y_{1i} - \mu(x_i))^2. \]  

Figure 2. The randomization distribution of Lin’s “interactions” estimator (left), the generalized Oaxaca-Blinder estimator based on “debiasing” the predictions \( \exp(\hat{\theta}_1^\top x_i) \) (center), and the generalized Oaxaca-Blinder estimator based on using the predictions \( \exp(\hat{\theta}_1^\top x_i) \) as a covariate in a linear model (right). All models control for state population, average miles per driver, and per capita income. The covariates are log transformed in the right two panels.

### 4.4. Isotonic Regression

Our last example is nonparametric. Suppose that the covariate dimension \( d \) is equal to one, and the outcome \( y_{1i} \) is bounded in the interval \([a, b]\). When the relationship between \( x_i \) and \( y_{1i} \) is expected to be monotone increasing, isotonic regression is a common modeling strategy. This method finds a function \( \hat{\mu}_1 \) by solving

\[ \hat{\mu}_1 \in \arg\min_{\mu \in \mathcal{M}} \frac{1}{n_1} \sum_{i=1}^{n_1} (y_{1i} - \mu(x_i))^2. \]  

where \( \mathcal{M} = \{ \mu : \mathbb{R} \to [a, b] \text{ nondecreasing} \} \).

That being said, we have found that using the in-sample prediction error as an estimate of MSE_{\text{in}}(\hat{\mu}) performs poorly with isotonic regression, due to overfitting. In a simple synthetic-data example shown in Figure 3, about 600 samples were needed per treatment arm before 95% confidence intervals achieved > 93% coverage. One explanation for why so many samples are needed is that the isotonic class is too flexible—see the jagged regression
function in Figure 3. The smoothed variants of isotonic regression that have been proposed in the literature (Jiang et al. 2011) may perform better in smaller samples. At the suggestion of a referee, we also explored estimating MSE at the suggestion of a referee, we also explored estimating MSE may perform better in smaller samples. At the suggestion of a referee, we also explored estimating MSE.

The parametric examples presented in Section 4 can all be extended to the “population” M-estimation problem (25).

5. Extension: A Generic Recipe for Parametric Models

The parametric examples presented in Section 4 can all be proved by translating standard arguments from the theory of M-estimation into the language of finite populations. This section gives a general result that lightens the effort of performing that translation in many cases. The result applies to parametric models where the parameter \( \hat{\theta} \) is estimated by solving a convex M-estimation problem.

Assumption A1. (Convex loss function)
Assume that \( \hat{\theta}_n \) is the solution to the following (random) optimization problem (25).

\[
\hat{\theta}_n \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \hat{\mathcal{L}}_n(\theta) := \frac{1}{n_1} \sum_{i=1}^{n} \ell(\theta, x_i, y_{1i}) \right\}, \tag{25}
\]

where \( \ell(\theta, x, y) \) is a loss function that is convex and twice-differentiable in its first argument.

The next assumption asks for the existence of a stable solution to the “population” version of Equation (25).

Assumption A2. (Existence and stability of population optima)
Assume that, for all large \( n \), there exists a vector \( \theta_n^* \) solving the “population” M-estimation problem (26).

\[
\theta_n^* = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \mathcal{L}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, x_i, y_{1i}) \right\}. \tag{26}
\]

Furthermore, suppose that for some \( c > 0 \), we have \( \nabla^2 \mathcal{L}_n(\theta) \geq d d \) whenever \( ||\theta - \theta_n^*|| \leq c \).

The assumption \( \nabla^2 \mathcal{L}_n(\theta) \geq d d \) near \( \theta_n^* \) guarantees the existence of a stable solution to \( \theta_n^* \), that is, \( \mathcal{L}_n(\theta) \) is smooth near its minimum.

Assumption A3. (Smooth loss function)
Assume that for all large \( n \), there exists a vector \( \theta_n^* \) solving the “population” M-estimation problem (26).

\[
\theta_n^* = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \mathcal{L}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, x_i, y_{1i}) \right\}. \tag{26}
\]

Furthermore, suppose that for some \( c > 0 \), we have \( \nabla^2 \mathcal{L}_n(\theta) \geq d d \) whenever \( ||\theta - \theta_n^*|| \leq c \).

The assumption \( \nabla^2 \mathcal{L}_n(\theta) \geq d d \) near \( \theta_n^* \) guarantees the existence of a stable solution to \( \theta_n^* \), that is, \( \mathcal{L}_n(\theta) \) is smooth near its minimum.

Assumption A3. (Smooth loss function)
Assume that for all large \( n \), there exists a vector \( \theta_n^* \) solving the “population” M-estimation problem (26).

\[
\theta_n^* = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \mathcal{L}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, x_i, y_{1i}) \right\}. \tag{26}
\]

Under the above assumptions, \( ||\hat{\theta}_n - \theta_n^*|| \) tends to zero in probability. If the map \( \theta \mapsto \mu_{\theta} \) is also smooth near \( \theta_n^* \), then that implies stability and bounded entropy.

Theorem 9. (General results for parametric models)
Assume that A1–A3 are satisfied. Then \( ||\hat{\theta}_n - \theta_n^*|| \to_p 0 \). If, in addition, \( ||\mu_\theta - \mu_{\phi}||_n \leq C||\theta - \phi|| \) for all \( \theta, \phi \in \mathbb{R}_c(\theta_n^*) \), then the sequence \( \{\mu_{\hat{\theta}_n}\}_{n \geq 1} \) is stable and typically simple.

6. Conclusion

In this article, we introduced an intuitive approach to performing covariate adjustment in randomized experiments. It can be summarized in a single sentence: “fill in the missing outcomes with an unbiased prediction model.” From a theoretical perspective, our main idea is that a little randomization goes a long way. As long as treatment assignments are randomized, then tools from empirical process theory can be applied even if all other quantities are nonrandom.
Many open questions remain. Perhaps, the most important one is whether the bounded entropy assumption can be removed, thereby opening the door to a much larger class of adjustment models. If one is willing to assume that $(X_i, Y_0, Y_1, Z_i)$ are iid samples from an infinite superpopulation, then clever use of sample splitting can circumvent these assumptions (Wager et al. 2016; Chernozhukov et al. 2018). Are similar results available in Neyman’s finite-population model? Recent work (Wu and Gagnon-Bartsch 2018) suggests that the answer might be “yes,” but establishing the stability condition for these complex models remains challenging. For example, Breidt and Opsomer (2000) studied the asymptotic properties of kernel regression in survey sampling settings (where design-based inference is the norm), but used simplifying “iid” assumptions in their analysis. They indicate that obtaining similar results under fixed-population assumptions would be substantially more challenging. We consider this to be an exciting direction for future work.

**Supplementary material**

The proofs of all the theorems and propositions, as well as additional numerical examples, can be found in the online supplement.

**Acknowledgments**

We thank Peng Ding, Avi Feller, Winston Lin, Art Owen and Fredrik Sajvé, as well as the participants in the causal inference reading group at UC Berkeley, for thoughtful comments and suggestions. We also thank the associate editor and two anonymous referees.

**ORCID**

Guillaume Basse © http://orcid.org/0000-0001-6032-5999

**References**

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020), “Sampling-Based Versus Design-Based Uncertainty in Regression Analysis,” *Econometrica*, 88, 265–296. [524,525]

Abadie, A., and Imbens, G. W. (2006), “Large Sample Properties of Matching Estimators for Average Treatment Effects,” *Econometrica*, 74, 235–267. [526]

Agresti, A. (1990), *Categorical Data Analysis*, New York: A Wiley-Interscience Publication. [532]

Albert, A., and Anderson, J. A. (1984), “On the Existence of Maximum Likelihood Estimates in Logistic Regression Models,” *Biometrika*, 71, 1–10. [531]

Athey, S., and Imbens, G. (2017), “Chapter 3—the Econometrics of Randomized Experiments,” in *Handbook of Field Experiments*, eds. A. V. Banerjee and E. Duflo, *Handbook of Economic Field Experiments*, Vol. 1, pp. 73–140. North-Holland. [525]

Barlow, R. E., and Brunk, H. D. (1972), “The Isotonic Regression Problem and Its Dual,” *Journal of the American Statistical Association*, 67, 140–147. [533]

Bartlett, J. W. (2018), “Covariate Adjustment and Estimation of Mean Response in Randomized Trials,” *Pharmaceutical Statistics*, 17, 648–666. [526]

Bauer, T. K., and Sinning, M. (2008), “An Extension of the Blinder–Oaxaca Decomposition to Nonlinear Models,” *ASA Advances in Statistical Analysis*, 92, 197–206. [526]

Belson, W. A. (1956), “A Technique for Studying the Effects of a Television Broadcast,” *Journal of the Royal Statistical Society*, Series C, 5, 195–202. [526]

Binder, D. A. (1983), “On the Variances of Asymptotically Normal Estimators From Complex Surveys,” *International Statistical Review*, 51, 279–292. [531]

Blinder, A. S. (1973), “Wage Discrimination: Reduced Form and Structural Estimates,” *The Journal of Human Resources*, 8, 436–455. [525,526]

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S., and Yu, B. (2016), “Lasso Adjustments of Treatment Effect Estimates in Randomized Experiments,” *Proceedings of the National Academy of Sciences*, 113, 7383–7390. [524,525,527,529,530]

Breidt, F. J., and Opsomer, J. D. (2000), “Local Polynomial Regression Estimators in Survey Sampling,” *The Annals of Statistics*, 28, 1026–1053. [535]

——— (2017), “Model-Assisted Survey Estimation With Modern Prediction Techniques,” *Statistical Science*, 32, 190–205. [526]

Bühlmann, P., and Yu, B. (2002), “Analyzing Bagging,” *Annals of Statistics*, 30, 927–961. [529]

Cao, W., Tsitsis, A. A., and Davidian, M. (2009), “Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean With Incomplete Data,” *Biometrika*, 96, 723–734. [525]

Cassell, C. M., Sarndal, C. E., and Wretman, J. H. (1976), “Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations,” *Biometrika*, 63, 615–620. [526]

Chakraborty, A., and Cai, T. (2018), “Efficient and Adaptive Linear Regression in Semi-Supervised Settings,” *The Annals of Statistics*, 46, 1541–1572. [526]

Chernozhukov, V., Chetverikov, D., Demirer, M., Dufo, E., Hansen, C., Newey, W., and Robins, J. (2018), “Double/Debiased Machine Learning for Treatment and Structural Parameters,” *The Econometrics Journal*, 21, C1–C68. [529,535]

Chung, E., and Romano, J. P. (2013), “Exact and Asymptotically Robust Permutation Tests,” *The Annals of Statistics*, 41, 484–507. [531]

Cohen, P. L., and Fogarty, C. B. (2020), “No-Harm Calibration for Generalized Oaxaca-Blinder Estimators,” arXiv:2012.09246. [530]

Ding, P., and Li, F. (2018), “Causal Inference: A Missing Data Perspective,” *Statistical Science*, 33, 214–237. [525,531]

Fairlie, R. W. (1999), “The Absence of the African-American Owned Business: An Analysis of the Dynamics of Self-Employment,” *Journal of Labor Economics*, 17, 80–108. [526]

Ferreira, J. A. (2015), “Some Models and Methods for the Analysis of Observational Data,” *Statistical Survey*, 9, 106–208. [526]

Fienberg, S. E., Tanur, J. M. (2018), “The Interlocking World of Surveys and Experiments,” *The Annals of Applied Statistics*, 12, 1157–1179. [525]

Firth, D., and Bennett, K. E. (1998), “Robust Models in Probability Sampling,” *Journal of the Royal Statistical Society*, Series B, 60, 3–21. [526,528,531]

Fogarty, C. B. (2018), “Regression-Assisted Inference for the Average Treatment Effect in Paired Experiments,” *Biometrika*, 105, 994–1000. [524,527]

Freedman, D. A. (2008a), “On Regression Adjustments to Experimental Data,” *Advances in Applied Mathematics*, 40, 180–193. [524,527]

——— (2008b), “On Regression Adjustments in Experiments With Several Treatments,” *Annals of Applied Statistics*, 2, 176–196. [524,527,528]

——— (2008c), “Randomization Does Not Justify Logistic Regression,” *Statistical Science*, 23, 237–249. [524,526,531]

Fuller, W. A. (2009), *Probability Sampling From a Finite Universe*, Chapter 1, pp. 1–93. Hoboken, NJ: Wiley. [531]

Guntuboyina, A., and Sen, B. (2018), “Nonparametric Shape-Restricted Regression,” *Statistical Science*, 33, 568–594. [533]

Hahn, J. (1998), “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, 66, 315–331. [526]

Hansen, B. B., and Bowers, J. (2009), “Attributing Effects to a Cluster-Randomized Get-Out-the-Vote Campaign,” *Journal of the American Statistical Association*, 104, 873–885. [525,526,531]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge, MA: Cambridge University Press. [525,531]

Jiang, X., Osl, M., Kim, J., and Ohno-Machado, L. (2011), “Smooth Isotonic Regression: A New Method to Calibrate Predictive Models,” *Journal of the American Statistical Association*, 106, 136–149. [533]
AMIA Summits on Translational Science Proceedings AMIA Summit on Translational Science 2011, pp. 16–20. [534]
Kang, J. D. Y., and Schafer, J. L. (2007), “Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean From Incomplete Data,” *Statistical Science*, 22, 523–539. [525]
Kennedy, E. H. (2016), *Semiparametric Theory and Empirical Processes in Causal Inference* (pp. 141–167). Cham: Springer International Publishing. [529]
Kline, P. (2011), “Oaxaca-Blinder as a Reweighting Estimator,” *American Economic Review*, 101, 532–37. [525,526]
Kosorok, M. (2008), *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics, Berlin: Springer. [529]
Lei, L., and Ding, P. (2020), “Regression Adjustment in Completely Randomized Experiments With a Diving Number of Covariates,” *Biometrika*. [524]
Li, X., and Ding, P. (2017), “General Forms of Finite Population Central Limit Theorems With Applications to Causal Inference,” *Journal of the American Statistical Association*, 112, 1759–1769. [525,530]
——— (2020), “Rerandomization and Regression Adjustment,” *Journal of the Royal Statistical Society*, Series B, 82, 241–268. [524,525,527,531]
Lin, W. (2013), “Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman’s Critique,” *Annals of Applied Statistics*, 7, 295–318. [524,525,527,530]
Liu, H., and Yang, Y. (2020), “Regression-Adjusted Average Treatment Effect Estimates in Stratified Randomized Experiments,” *Biometrika*, 107, 935–948. [524,527]
Lohr, S. (2009), *Sampling: Design and Analysis*, Advanced (Cengage Learning). Cengage Learning. [525]
Mukerjee, R., Dasgupta, T., and Rubin, D. B. (2018), “Using Standard Tools From Finite Population Sampling to Improve Causal Inference for Complex Experiments,” *Journal of the American Statistical Association*, 113, 868–881. [525]
Negi, A., and Wooldridge, J. M. (2020), “Robust and Efficient Estimation of Potential Outcome Means Under Random Assignment,” arXiv:2010.01800. [526]
Neyman, J. (1923), “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9,” *Statistical Science*, 5, 465–472. [524,525,527,530]
Oaxaca, R. (1973), “Male-Female Wage Differentials in Urban Labor Markets,” *International Economic Review*, 14, 693–709. [525,526,532]
Olsen, R., Orr, L., Bell, S., and Stuart, E. (2013), “External Validity in Policy Evaluations that Choose Sites Purposively,” *Journal of Policy Analysis and Management*, 32, 107–121. [524]
Peters, C. C. (1941), “A Method of Matching Groups for Experiment With no Loss of Population,” *The Journal of Educational Research*, 34, 606–612. [526]
Roberts, G., Rao, N. K., and Kumar, S. (1987), “Logistic Regression Analysis of Sample Survey Data,” *Biometrika*, 74, 1–12. [531]
Robins, J. (1986), “A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect,” *Mathematical Modelling*, 7, 1393–1512. [526]
Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994a), “Estimation of Regression Coefficients when some Regressors are not Always Observed,” *Journal of the American Statistical Association*, 89, 846–866. [525]
——— (1994b), “Estimation of Regression-Coefficients when some Regressors are not Always Observed,” *Journal of the American Statistical Association*, 89, 846–866. [526]
Rosenbaum, P. R. (2002), “Covariance Adjustment in Randomized Experiments and Observational Studies,” *Statistical Science*, 17, 286–327. [525]
Rosenberger, W. E., Uschner, D., and Wang, Y. (2019), “Randomization: The Forgotten Component of the Randomized Clinical Trial,” *Statistics in Medicine*, 38, 1–12. [524]
Rosenblum, M., and van der Laan, M. (2011), Robust Analysis of RCTs Using Generalized Linear Models. New York: Springer. pp. 187–199. [525,526]
Rothe, C. (2018), “Flexible Covariate Adjustments in Randomized Experiments,” Mannheim. Available at: https://madoc.pub.uni-mannheim.de/52249/. [529]
Rubin, D. (1974), “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66, 688–701. [525]
Ruhm, C. J. (1995), “Alcohol Policies and Highway Vehicle Fatalities,” Working Paper 5195, National Bureau of Economic Research. [527]
Särndal, C., Swensson, B., and Wretman, J. (2003), *Model Assisted Survey Sampling*. Springer Series in Statistics. New York: Springer. [525]
Schochet, P. Z. (2010), “Is Regression Adjustment Supported by the Neyman Model for Causal Inference?” *Journal of Statistical Planning and Inference*, 140, 246–259. [524]
Shen, E. (1940), “Experimental Design and Statistical Treatment in Educational Research,” *The Journal of Experimental Education*, 8, 346–353. [526]
Särndal, C.-E., and Wright, R. L. (1984), “Cosmetic Form of Estimators in Survey Sampling,” *Scandinavian Journal of Statistics*, 11, 146–156. [526,528]
Tibshirani, R. J., Hoelling, H., and Tibshirani, R. (2011), “Nearly-Isotonic Regression,” *Technometrics*, 53, 54–61. [534]
vanderLaan,M.,andRose,S.(2011),TargetedLearning:CausalInferenceforObservationalandExperimentalData, Springer Series in Statistics. New York: Springer. [532]
vanderVaart,A.,andWellner,J.(1996),WeakConvergenceandEmpiricalProcesses:WithApplicationstoStatistics, Springer Series in Statistics. New York: Springer. [529]
wagen,S.,Du,W.,Taylor,J.,andTibshirani,R. J. (2016), “High-Dimensional Regression Adjustments in Randomized Experiments,” *Proceedings of the National Academy of Sciences*, 113, 12673–12678. [535]
Wu, C., and Sitter, R. R. (2001), “A Model-Calibration Approach to Using Complete Auxiliary Information From Survey Data,” *Journal of the American Statistical Association*, 96, 185–193. [526]
Wu, E., and Gagnon-Bartsch, J. A. (2018), “The Loop Estimator: Adjusting for Covariates in Randomized Experiments,” *Evaluation Review*, 42, 458–488. [525,535]
Wu, L., and Ding, P. (2020), “Randomization Tests for Weak Null Hypotheses in Randomized Experiments,” *Journal of the American Statistical Association*, 1–16. [529]
Yang, L., and Tsai, A. A. (2001), “Efficiency Study of Estimators for a Treatment Effect in a Pretest–Posttest Trial,” *The American Statistician*, 55, 314–321. [524]
Zhao, A., and Ding, P. (2020), “Covariate-Adjusted Fisher Randomization Tests for the Average Treatment Effect,” arXiv:2010.14555. [531]