# Does Regression Produce Representative Estimates of Causal Effects?

**Peter M. Aronow** Yale University
**Cyrus Samii** New York University

*With an unrepresentative sample, the estimate of a causal effect may fail to characterize how effects operate in the population of interest. What is less well understood is that conventional estimation practices for observational studies may produce the same problem even with a representative sample. Causal effects estimated via multiple regression differentially weight each unit's contribution. The "effective sample" that regression uses to generate the estimate may bear little resemblance to the population of interest, and the results may be nonrepresentative in a manner similar to what quasi-experimental methods or experiments with convenience samples produce. There is no general external validity basis for preferring multiple regression on representative samples over quasi-experimental or experimental methods. We show how to estimate the "multiple regression weights" that allow one to study the effective sample. We discuss alternative approaches that, under certain conditions, recover representative average causal effects. The requisite conditions cannot always be met.*

Different people, communities, or countries respond in different ways to the interventions, natural events, or structural conditions that political scientists study. Some political science research has focused on describing effect heterogeneity. For example, in comparative politics, Hidalgo et al. (2010) find that the effects of economic shocks on conflict in Brazil are larger in areas with high levels of preexisting inequality. In American politics, Arceneaux and Nickerson (2009) find that the effects of canvassing on voter participation are larger for those nearer to the threshold of indifference over participating versus not participating. In international relations, Colgan (2010) finds that the effects of oil resources on a country's propensity to engage in conflict depend on the political conditions within that country. It would strain credulity to think that one could model *all* determinants of such heterogeneity. Researchers are typically limited in their ability to characterize such heterogeneity by a lack of data or knowledge about the functional forms that such heterogeneity obeys. Therefore, political scientists must be content with estimating causal effects that, to some degree, average over unmodeled heterogeneity. The "averages" that result from any analysis will be based on contributions from a set of units that may not resemble the general population of interest. Such effect heterogeneity is what prompts political scientists occasionally to question the "generalizability" or "external validity" of research findings.

External validity is a central concern in current debates about methods for causal inference. Following Campbell's (1957) original formulation, questions of external validity are ones that ask, "To what population, settings, and variables can this effect be generalized?" These are distinguished from questions of internal validity, which address questions of whether a causal effect is credibly estimated or, to quote Campbell again in the case of a proposed estimate of a causal effect, "Did in fact the experimental stimulus make some significant difference in this specific instance?" (297). Succinctly put, external validity concerns generalizing to other populations of interest beyond the one that was the basis of a given study (see, e.g., Morton and Williams 2010, 264–275).

Some methodologists praise experiments and natural experiments using instrumental variables and regression discontinuity designs for their strong internal validity (Angrist and Pischke 2009; Dunning 2008; Imbens and

Peter M. Aronow is Assistant Professor, Department of Political Science, Yale University, 77 Prospect St., New Haven, CT 06520 (peter.aronow@yale.edu). Cyrus Samii is Assistant Professor, Department of Politics, New York University, 19 West 4th St., New York, NY 10012 (cds2083@nyu.edu).

The authors thank Neal Beck, Allan Dafoe, Thad Dunning, Andrew Gelman, Don Green, Winston Lin, Vera Troeger, three *AJPS* reviewers, and the *AJPS* editor for helpful comments, and Alan Gerber, Nathan Jensen, and Gregory Huber for the data and replication code that we use in our illustrations. Replication materials are available in the AJPS Data Archive on Dataverse (http://dvn.iq.harvard.edu/dvn/dv/ajps).

***American Journal of Political Science,*** Vol. 60, No. 1, January 2016, Pp. 250–267



DOI: 10.1111/ajps.12185

Rubin 2011; Morgan and Winship 2007; Morton and Williams 2010; Robinson, McNulty, and Krasno 2009). They also praise matching methods as an alternative to regression because matching increases internal validity by relieving dependence on functional form restrictions (Ho et al. 2007). A common criticism of experiments, natural experiments, and matching is that such methods increase internal validity, but in doing so they introduce problems for external validity in that it is no longer clear whether the estimated effects generalize to a population of interest. Experiments are carried out using subpopulations that may differ markedly from broader populations for which we want to estimate effects, and instrumental variables methods and regression discontinuity designs identify "local" causal effects that apply only to highly specific subgroups (Deaton 2010; Heckman and Urzua 2010). Matching has us discard unmatched units and therefore mechanically limits the scope of our inference (Banerjee and Duflo 2009, 162–63).

Taking stock of these points, Dunning (2008, 291) warns that "as the plausibility of 'as if' random assignment increases, the population of units for which a causal effect may be reliably estimated may be quite small." As Campbell (1984) and Rosenbaum (1999) have noted, researchers have tended to interpret this fact in terms of a trade-off between internal and external validity. Researchers seem to believe that when pursuing empirical evidence about a causal relationship, they face a choice between (a) privileging internal validity by using experiments or hunting out idiosyncratic natural experiments, in which case they will likely have to accept working with highly specific subpopulations, and (b) minimizing external validity problems by working with more model-dependent regression methods that nonetheless "fully exploit" data sets that are more representative of the population of interest. This belief is succinctly conveyed in the following quote by an eminent economist, Pranab Bardhan, writing on the trend of using randomized controlled trials (RCTs) in development economics:

> RCTs face serious challenges to their generalizability or "external validity." Because an intervention is examined in a microcosm of a purposively selected population, and not usually in a randomly sampled population for any region, the results do not generalize beyond the boundaries of the study. For all their internal flaws, the older statistical studies, often based on regional samples, permitted more generalizable conclusions. Neither method has a monopoly on correctness. (Bardhan 2013)

In this article, we show that this perceived trade-off, and the choice that it suggests, is an illusion. Our argument can be summarized as follows. Suppose one has data from a sample that is representative of a population for which one wants to estimate a causal effect. Call this the *nominal* sample.[1] If one uses this sample to estimate a causal effect via a multiple regression model, the units in the nominal sample will contribute to the effect estimate to differing extents. We can derive weights that measure the contribution from each member of the nominal sample. By reweighting the nominal sample with these *multiple regression weights*, we characterize the *effective sample* that multiple regression actually uses to estimate the effect. The effective sample will, generally speaking, differ from the nominal sample. Even though one starts with a sample representative of the population of interest, what one obtains is an effect for which there are still reasons to question generalizability to the population of interest. External validity problems have not been avoided.

Our focus on multiple regression is motivated by the fact that it remains as a workhorse approach for estimating causal effects in political science. Some researchers who conduct regression studies may not think that the trade-off described above is relevant to them. Even so, it is crucial for such researchers to understand the external validity issues that arise in regression studies. We develop this point further in the conclusion. Our formal analysis builds on results of Angrist and Krueger (1999, 1311–12) and Angrist and Pischke (2009, chap. 3), who show that multiple regression estimates are equivalent to weighted averages of unit-specific contributions, with the resulting multiple regression weights driven by the conditional variance of the causal factor of interest. Most textbook discussions of one-way fixed-effects regressions, for example, emphasize how such models rely only on "within" variation, effectively excluding from the calculation of effects any groups for which there is no within variation in the causal factor (Beck and Katz 2001). But such discussions paint too black-and-white a picture: among cases exhibiting within variation, some cases still may be given much more weight than others. This is what our results clarify. Regression textbooks typically use the "leverage" statistic to measure how much a given unit influences a regression estimate (Davidson and MacKinnon 2004, 77–79). As we show below, the leverage does not allow us to recover a unit's contribution to a causal effect estimate; rather, the leverage obscures this contribution by

[1] By *nominal* we mean the dictionary definition—"in name only." We do not mean to refer to *nominal data* which is often used to mean "categorical data."

combining influence due to the distribution over both the treatment variables and control variables. Other related results are given in Humphreys (2009) and Imai and Kim (2012). The results in this article complement the points raised by King and Zeng (2006) in their discussion of the role of covariate overlap in robust estimation of regression parameters and by Imai, King, and Stuart (2008) on how sample selection and biased estimation combine to bias a sample estimate relative to a population target.

We extend Angrist and Krueger's (1999) and Angrist and Pischke's (2009) results in two ways. First, we relate them to a more general model of "potential outcomes," which is the dominant analytical framework for studying causal effects in the social sciences. Angrist and Krueger (1999) characterize a generic linear regression estimate; bringing in potential outcomes allows us to express these results in terms of causal quantities, clarifying implications for causal inference. Angrist and Pischke (2009) develop results in the case of a binary treatment variable and a saturated dummy variable regression; we allow for arbitrary types of treatment variables and control variable specifications, given that the vast majority of regression studies use specifications more complicated than binary treatment with dummy variable controls. Second, we show how to use multiple regression weights to characterize the effective sample that multiple regression uses to estimate a causal effect. We use two recent political science studies to illustrate how researchers can study their effective sample and thereby evaluate the generalizability of their findings through visualization and weighted summary statistics. An appendix demonstrates that the results we obtain for linear regression also carry through to generalized linear models (logit, probit, etc.). We explain how simple extensions of multiple regression, such as random coefficient models, do not resolve the problem. We then discuss approaches that actually do, under certain restrictive conditions, recover the average causal effect from a representative sample.

## Data, Potential Outcomes, and Causal Effects

Our analytical framework allows for causal effects to vary over units in the population. In addition, we presume that the analyst is limited in her ability to characterize such effect variation, whether because some determinants of such variation are unmeasured or the correct specification is unknown. Given the types of causal phenomena that political scientists study, this is the typical scenario for empirical research, and these two conditions establish external validity as a relevant concern.

Suppose that we begin with a population of interest and draw a sample indexed by $i = 1, \ldots, n$. We assume that the sample is self-weighting and representative—each member of the population has an equal probability of being selected. Such a sample requires no special adjustment (e.g., via weighting) to be made representative of the population.[2] In some applications, it does not make sense to think in terms of sampling from a population. An example is international relations studies with data sets that contain all countries in the world. In these cases, one may assume that the conditions that we see in the countries in our data set are but one of the set of possible realizations that could have materialized given that nature is stochastic. This latter formulation is known in the methodology literature as a "super-population" assumption. All of the results in this article continue to apply if we substitute sampling from a fixed population with the super-population assumption.[3] For all $i$ that appear in the sample, we measure (a) a causal factor of interest, or "treatment," $D_i$, which may be continuous or binary; (b) a $K$-length row vector of control variables, $X_i$, that exhibit no perfect colinearities or measurement error; and (c) an outcome, $Y_i$. The sampling or super-population assumption implies that $D_i$, $X_i$, and $Y_i$ are random variables, although not necessarily independent.

Our goal is to measure the causal effect of $D_i$ on $Y_i$. In the current social science literature, a common and extremely clarifying approach to analyzing causal effects is to use "potential outcomes" (Holland 1986; Neyman[1923] 1990; Rubin 1978). The idea is very simple: for every potential value of $D_i$, each unit $i$ is assumed to possess a well-defined response. Thus, if $D_i$ can take on the values $d_1$, $d_2$, ..., then unit $i$'s potential outcomes are given by the corresponding values $Y_i(d_1)$, $Y_i(d_2)$, .... Causal effects for unit $i$ are formalized as comparisons (e.g., differences or ratios) between potential outcomes. The average causal effect for a population is the average of these unit-level causal effects for units in this population. The potential outcomes framework is highly general, and even researchers who operate within the conventional

[2] To generalize the results below to the case where sample weights or some other weighting scheme must be applied to make the sample representative of the target population, one merely needs to carry those weights through all of the summations that appear in the mathematical analysis. No other complications arise.

[3] Abadie et al. (2010) discuss how the usual frequentist statistical inference is valid even without a super-population assumption, given stochasticity in the treatment. Berk, Western, and Weiss (1995) and Gill (2001) contrast such frequentist constructions with Bayesian inferential frameworks.

regression-based framework "are working within a [potential outcomes] basis either implicitly or explicitly" (Morton and Williams 2010, 109).[4]

Our primary interest in this article is to study the representativeness of linear regression estimators. To this end, we impose assumptions on the underlying causal model that satisfy four criteria: (a) they are as generous to multiple regression as possible, (b) they establish situations in which external validity concerns are present, (c) they stay close to assumptions that are implicit in current practice, and (d) they can in principle be satisfied exactly or approximately in a typical regression analysis. We could use more general assumptions, but this would only make regression look worse by combining issues of primary concern here (generalizability and external validity) with other issues (inadequate control of confounders and therefore problems of internal validity). The only way that our assumptions could be more restrictive would be to assume either that causal effects are constant for all units or that all sources of effect heterogeneity have been modeled. Such additional assumptions are very heroic and most likely unrealistic, but if they do hold, there would be no reason to worry about the types of external validity problems that we discuss here.

Based on these four criteria, we suppose that potential outcomes take the following form:

$$Y_i(d) = Y_i(0) + \tau_i \cdot d, \tag{1}$$

where $\tau_i$ is the causal effect for unit $i$ and $d$ is an arbitrary value of $D_i$. When the treatment is binary, $D_i \in \{0, 1\}$, then this characterization is completely general. In that case, when $D_i = 0$, we observe the realization of the potential outcome under the "control condition": $Y_i = Y_i(0) + \tau_i \cdot 0 = Y_i(0)$. When $D_i = 1$, we observe the realization under the "treated condition": $Y_i = Y_i(0) + \tau_i = Y_i(1)$. The causal effect for unit $i$ is well defined as the difference between potential outcomes: $\tau_i = Y_i(1) - Y_i(0)$. The sample average causal effect is the average of these unit-level causal effects for units in the sample. When $D_i$ is continuous, Expression (1) is valid when "local linearity" holds—that is, while causal effects may be arbitrarily heterogeneous across units, such effects are linear in $D_i$ for each unit.[5]

We allow for the control variables in $X_i$ to be temporally predetermined relative to $D_i$ and the baseline[6] value, $Y_i(0)$, but potentially correlated with both $D_i$ and $Y_i(0)$.[7] However, we assume that the control variables in $X_i$ are indeed sufficient for removing any confounding in the relationship between $D_i$ on $Y_i$, so that

$$(Y_i(0), \tau_i) \perp\!\!\!\perp D_i | X_i. \tag{2}$$

This expression is read as "restricting ourselves to units with the same $X_i$, the treatment $D_i$ is assigned to such units in a manner that is independent of the baseline potential outcome $Y_i(0)$ and the treatment effect $\tau_i$." We also assume that the conditional expected value of $D_i$ can be represented as a linear combination of the covariates in $X_i$. By *linearity* here, we mean linearity in parameters, in that $X_i$ may contain higher-order terms, interactions, and dummy variables based on coarsened covariates (Wooldridge 2009, 46). So this linearity condition is not so restrictive when one considers the possibility of enriching one's specification for $X_i$ in these ways.[8] Those using multiple regression should want it to hold. Formally, this assumption may be written as

$$\mathrm{E}[D_i | X_i] = V_i \omega, \tag{3}$$

where $V_i = (1 \quad X_i)$, $\omega$ is a $K + 1$ column vector of coefficients, and $\mathrm{E}[.]$ averages over the units in the population.[9]

Readers should avoid misunderstanding the point of assuming linearity in Expressions (1) and (3). It provides a very favorable scenario for researchers who want to use multiple regression to estimate causal effects. The linearity restrictions allow us to focus precisely on the issue at hand—questions of external validity and the generalizability of linear regression estimators. Working with

[4] Morton and Williams (2010) use the term *Rubin Causal Model* (RCM) in place of potential outcomes to acknowledge the seminal contributions of statistician Donald B. Rubin (as in Rubin 1978).

[5] Note that this model embeds the so-called stable unit treatment value assumption, which states that the potential outcome associated with unit $i$ is determined solely by unit $i$'s treatment value, $D_i$ (Rubin 1980). This rules out interference effects or incomplete measurement of treatment conditions.

[6] By *baseline value* we mean to suggest "a value that serves as a basis for comparison," rather than to suggest outcomes that occur temporally prior to the application of a treatment. *Comparator value* may be a clearer term, though not one that is as familiar.

[7] Because $X_i$ is assumed to be temporally predetermined relative to $D_i$, there is no risk of introducing posttreatment bias by controlling for $X_i$ (King and Zeng 2006; Rosenbaum 1984). Technically, this condition is not even necessary—as in the case of posttreatment variables that are nonetheless strictly exogenous (Chamberlain 1984). What is necessary is that the conditional independence assumption in Expression (2) holds.

[8] Angrist and Krueger's (1999) explain the consequences of applying a linear model when effects are not linear, although like us they assume the treatment is linear in the covariates. See Angrist and Krueger's (1999, 1311–12) for a discussion.

[9] For those who prefer Pearl's (2009) analytical framework for causality, Expression (2) establishes the relevant family of graphs, Expressions (1) and (3) establish the relevant structural equations, and our interest is in effects of manipulations on $d$ in Expression (1).

more general types of potential outcomes and confounding would require that we also deal with questions of internal validity and regression model misspecification. If the linearity assumptions were not valid, then the consequences for multiple regression would only be worse than what we obtain below.

The goal is to estimate the average causal effect for the population. This is simply the expected value of the unit-level causal effects:

$$\bar{\tau} = \mathrm{E}\,[\tau_i] = \mathrm{E}\left[\frac{Y_i(d') - Y_i(d)}{d' - d}\right], \tag{4}$$

where $d' \neq d$ are two arbitrary values that are possible for $D_i$.[10] The population average causal effect is the standard inferential target for social science researchers (Angrist and Pischke 2009; Imai, King, and Stuart 2008), although we acknowledge that there may be cases where other features of the population effect distribution are of interest (e.g., quantile effects or ratio effects). Given that potential outcomes are determined according to Expression (1) and our inferential target is $\bar{\tau}$, what does a regression of $Y_i$ on $D_i$ and covariates estimate?

## Multiple Regression, Causal Effects, and Effective Samples

By combining a few well-established results in linear regression theory, we can determine the *effective sample* that contributes to the estimation of a causal effect via regression. The effective sample takes into account the *multiple regression weight* given to each sample member in constructing the estimate. For example, if one begins with a sample of 50% women and 50% men, but then women are given three times as much weight as men in producing the estimate, the effective sample would be 75% women and 25% men. We develop these results in the context of linear regression model fit via ordinary least squares (OLS). What we refer to as "multiple regression" is a regression of the scalar outcome, $Y_i$, on a set of regressors—in this case, the scalar treatment, $D_i$, and the control vector, $X_i$. We explain how the results apply to nonlinear models fit via maximum likelihood and random coefficient models.

[10]The assumption of linear potential outcomes in the treatment means that $\bar{\tau}$ does not depend on what we choose for $d'$ and $d$. When we discuss more general models for causal effects below, this scaling becomes more consequential.

### Ordinary Least Squares Regression

As an analogy to the causal relationship given in Equation (1), we might consider estimating the linear regression model,

$$Y_i = \gamma + \lambda D_i + \eta_i, \tag{5}$$

where $\gamma$ is an intercept and $\lambda$ a slope coefficient to be estimated, while $\eta_i$ is an error term. However, we have presumed that $D_i$ and $Y_i$ may be jointly correlated with the control variables in $X_i$. So, the exclusion of the $X_i$ variables from the specification will give rise to omitted variable bias if we fit Equation (5) via ordinary least squares to estimate the partial relationship between $D_i$ and $Y_i$.

The concern about omitted variable bias leads us to consider the following linear regression model:

$$Y_i = \alpha + \beta D_i + X_i\gamma + \epsilon_i, \tag{6}$$

where $\alpha$, $\beta$, and $\gamma$ are model coefficients and $\epsilon_i$ is an error term. The regression specification in Expression (6) is a version of the workhorse model used by political scientists (and researchers in many other disciplines) for estimating causal effects. Such a model assigns a constant effect, $\beta$, for each unit, with the error term, $\epsilon_i$, standing in for all unit-specific idiosyncrasies. Classical treatments of linear regression (as in, e.g., Greene 2008 or Wooldridge 2009) work on the basis of this constant effects assumption. Of course, to do so is to sweep under the rug the heterogeneity that gives rise to concerns about representativeness or external validity. The goal here is to clarify the implications of the effect heterogeneity that we know to be relevant in our empirical analyses.

We fit Model (6) using OLS to obtain estimates of the parameters, $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$. The estimate $\hat{\beta}$ is our regression estimate of the causal effect of $D_i$. However, fitting Model (6) via OLS induces a weighting scheme that implies that contributions from sample members are used differentially. The weights can be characterized completely via the regressors and do not depend on the outcome. Applying results from partial regression (Greene 2008, 27–29) along with Conditions (1) and (3), multiple regression generates a weighted average of causal effects of the form

$$\hat{\beta} \xrightarrow{p} \frac{\mathrm{E}\,[w_i\tau_i]}{\mathrm{E}\,[w_i]}, \text{ where } w_i = (D_i - \mathrm{E}\,[D_i|X_i])^2. \tag{7}$$

We call $w_i$ the *multiple regression weight* for unit $i$. The relation $\xrightarrow{p}$ refers to convergence in probability as the sample size, $n$, gets larger and larger (i.e., the approximation to $\hat{\beta}$ given by the right-hand side becomes exact as the sample size grows). The derivation is given in

Theorem A.1 in Appendix A. By the definition of conditional variance,

$$\mathrm{E}[w_i|X_i] = \mathrm{Var}[D_i|X_i]. \quad (8)$$

A simple way to interpret the result is that more weight goes to units whose treatment values ($D_i$) are not well explained by the covariates ($X_i$).[11] In the case of a binary treatment, Expression (8) reduces to a simple function of the propensity score ($\mathrm{E}[w_i|X_i] = \pi(X_i)[1 - \pi(X_i)]$, where $\pi(X_i) = \mathrm{E}[D_i|X_i]$, the propensity score). Angrist and Pischke (2009, 69–80) show that with a binary treatment and a saturated dummy variable control vector, the differences between $\hat{\beta}$ and the "average effect of the treatment on the treated" are captured precisely by this difference between the multiple regression weight and the propensity score. Furthermore, with a binary treatment, inverse propensity score weighting would break the dependence between $D_i$ and $X_i$ and therefore allows one to sidestep issues created by the multiple regression weights. Indeed, if inverse propensity score weighting is applied and there is no misspecification for the propensity scores, the regression weights are all, asymptotically, unity within treatment strata. This anticipates the discussion of methods for estimating representative causal effects, which we take up below.

The result may seem to resemble the expression for the "leverage" statistic, but the leverage, as defined by Davidson and MacKinnon (2004, 77–79), measures unit $i$'s overall distance from the center of the data in terms of all of the predictors, $D_i$ and $X_i$, and therefore characterizes the effect of dropping $i$ on the overall coefficient vector ($\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$). By contrast, the multiple regression weight $w_i$ measures only the contribution of unit $i$'s effect, $\tau_i$, to the construction of $\hat{\beta}$. In fact, units with high leverage may have low multiple regression weights, and vice versa. (Appendix B presents some clarifying examples.) It is precisely $w_i$ that is the appropriate measure of the extent to which $\hat{\beta}$ incorporates the effect associated with unit $i$.

[11]Expression (8) implies that there exist simple tests of the null hypothesis that the covariate distributions of the nominal and effective samples are equivalent. For example, a heteroskedasticity test along the lines of Breusch and Pagan (1979) for a regression of the $D_i$ values on the $X_i$ values would test against this null. In our view, such tests are of little practical value because they do little to expose how the effective sample departs from the nominal sample. While both examples we present in the fourth section will yield rejection of the null hypothesis using such tests (with $p < .001$), we will see—using more appropriate visualization techniques—that these tests are not necessarily prognostic of any substantively meaningful differences between the effective and nominal samples.

## Multiple Treatments in One Regression

Our analysis has focused on the case of one treatment and a vector of control variables. If one is estimating causal effects for multiple treatments in a single regression model, then one should compute a separate set of multiple regression weights for each treatment. Effectively, one takes each treatment one-by-one to be $D_i$ in the expressions above, and for the purposes of computing the multiple regression weight for that treatment, considers all other treatments as components of $X_i$. It is not generally the case that a unit's weight will be the same for each of the treatments. (The leverage would obscure this fact.)

# The Effective Sample from Ordinary Least Squares Regression

One may use the multiple regression weights (and *not* the leverage) to characterize the *effective sample* that gives rise to a causal effect estimate. Knowing the effective sample is what allows one to relate effect estimates to units' background characteristics and to evaluate for what kind of population the estimated effects are relevant. For example, suppose you have reason to suspect that the effects of a treatment vary on the basis of certain background characteristics. It is precisely the effective sample that you need in order to compute summary statistics for such background characteristics and to see whether the effective sample differs from the population on the basis of these characteristics.[12]

Expression (7) implies that the effective sample is given by reweighting the original sample by the multiple regression weights—the $w_i$s. We may estimate each of the multiple regression weights with the estimator $\hat{w}_i = \tilde{D}_i^2$, where $\tilde{D}_i$ is the residual from a regression of $D_i$ on $X_i$. Then, much as we compute summary statistics for the nominal sample, we may use the multiple regression weights to estimate summary statistics for the effective sample. We can estimate the mean of a covariate $Z_i$ in the effective sample with

$$\hat{\mu}(Z_i) = \frac{\sum_{i=1}^{n} \hat{w}_i Z_i}{\sum_{i=1}^{n} \hat{w}_i} \xrightarrow{p} \frac{\mathrm{E}[w_i Z_i]}{\mathrm{E}[w_i]} = \mu(Z_i), \quad (9)$$

[12]Background characteristics that modify the effect of a treatment are known in the methodology literature as "moderators" (Barron and Kenny 1986). In other words, they are covariates for which there are nonzero interaction effects with the causal factor. VanderWeele and Robins (2007) discuss challenges for causal inference when studying moderators (which they term "effect modifiers").

where $\mu(.)$ refers to the mean for the effective sample.[13] For example, we could define $Z_i$ as an indicator variable for "unit is in Europe" to get the share of units in the effective sample that are in Europe. In panel data sets (assuming that the number of time periods is large), we can characterize how heavily weighted each cross-sectional unit (e.g., country, individual, or dyad) is in the effective sample. Similar calculations—redefining $Z_i$ appropriately—permit consistent estimation of each of the covariate cumulative density functions in the effective sample.[14]

In a world of effects that are heterogeneous, characterizing the covariate profile for the effective sample is crucial. We learn about the type of population to which our effect estimates apply. The multiple regression weights provide the basis for doing this. When some types of observations have multiple regression weights of zero, this means that the covariates *completely* explain their treatment condition. Such units do not contribute to the estimate that we obtain from the study, and therefore there is no statistical basis to conclude that the results of the study are applicable to them. When some types of observations have multiple regression weights that are very small relative to other observations, it means that they are poorly represented in the effective sample from which the effect estimate is obtained. Without further assumptions, we have a weak statistical basis for claiming that our effect estimate is applicable to them, as their contribution is overwhelmed by the contribution of those with larger weights.

### Randomized Experiments

We had assumed above that $X_i$ confounded the relationship between $D_i$ and $Y_i$, but it is instructive to stop and think for a moment what would happen were this not the case. That is, suppose that $D_i$ were randomly assigned, as would be the case in a well-executed randomized experiment with equal assignment probabilities. Then $D_i$ would be independent of $\tau_i$ as well as $X_i$, in which case $w_i$ would be independent of $\tau_i$. As a result, the weights in Expression (7) would effectively cancel out, and $\hat{\beta}$ would converge to $\bar{\tau}$.[15] (See Corollary A.1 in the appendix.) While controlling for $X_i$ was not necessary to deal with confounding, doing so could help to make the estimate of $\hat{\beta}$ more precise. This is how the marriage of random sampling, random assignment, and regression can work to produce internally valid, representative, and precise estimates of causal effects (Lin 2013).[16]

Randomized experiments are rare in political science, which is the reason that we consider methods of covariate control to eliminate confounding. The implicit weighting of multiple regression can lead us away from our target, $\bar{\tau}$. In particular, it is the potential for association between $w_i$ and $\tau_i$ that causes the distortion. The regression estimate, $\hat{\beta}$, may incorporate contributions from sample members in a way that departs substantially from equal representation among the population of interest. The effective sample may be a gross distortion of the type of population that one intends to characterize.

## Other Regression-Based Estimators

Here we consider implications of our result for other regression-based estimators, namely, estimates from nonlinear maximum likelihood estimators and from random coefficient models. We show that these estimators apply the same multiple regression weights that OLS does, either exactly or approximately.

### Maximum Likelihood Estimation and Generalized Linear Models

The results for linear regression carry over directly to nonlinear regression models fit via maximum likelihood. This includes the class of generalized linear models, such as logit, probit, multinomial logit or probit, Poisson regression, and parametric or semiparametric (e.g., Cox)

[13] As sample estimates, the set of $\hat{w}_i$s would vary from sample to sample, and therefore the effective sample summary statistics computed with them—the $\hat{\mu}(Z_i)$s—have sampling distributions. The associated standard errors can be approximated with bootstrap resampling. Alternatively, one could use a linearized approximation to the variance as is usually done for weighted means with random weights and other ratio estimators (Särndal, Swensson, and Wretman 1992, 172–81).

[14] When one is estimating effects for multiple treatments in one regression, the effective samples for each treatment may indeed differ.

[15] Some experimental designs assign values of $D_i$ to different units with varying probabilities, including many block-randomized experiments. For such designs, it is necessary to include a statistical adjustment for the unequal probabilities associated with the experimental design. A common approach is to use one-way fixed effects to account for the blocking strata, but such a regression will estimate a reweighted causal effect as in Expression (7). Thus, while a one-way fixed-effects regression of this sort is neither unbiased nor consistent for $\bar{\tau}$, unbiased estimators do exist for $\bar{\tau}$ (Aronow and Middleton 2013; Gerber and Green 2012, 116–21).

[16] In this way, it is not multiple regression estimation per se that is a problem, but rather the false (implicit) assumption that multiple regression estimation implies equal use of all sample contributions to a causal effect estimate in cases when $D_i$ is not randomized.

duration models (Long 1997; McCullagh and Nelder 1999). For example, consider a logistic regression that is analogous to the linear regression that we developed above:

$$\Pr[Y_1 = 1] = \text{logit}^{-1}(\alpha_\ell + \beta_\ell D_i + X_i\gamma_\ell),$$

where the subscripting by $\ell$ indicates that the coefficients are for a logistic regression model, and $\text{logit}^{-1}(a) = 1/[1 + \exp(-a)]$ is the inverse of the logit link function (Long 1997, chap. 3). Typically, we estimate the coefficients in such a model through maximum likelihood, obtaining $(\hat{\alpha}_\ell \quad \hat{\beta}_\ell \quad \hat{\gamma}_\ell)'$. For a nonlinear model such as this, there is no closed-form solution for the weights that allow us to write $\hat{\beta}_\ell$ as a weighted average of unit-level contributions. We can obtain an approximation of arbitrary precision for the weights by using a Taylor expansion of the inverse link function, $\text{logit}^{-1}(.)$. Suppose we center such an expansion on the logit of the mean of $Y_i$. Then, to a first approximation, the only difference between $\hat{\beta}_\ell$ and what we would get were we to just use OLS on these data is a scaling factor that depends solely on the mean of $Y_i$. Therefore, the weights implied by a logistic regression are approximately equal to the multiple regression weights from OLS. As with OLS, the amount of weight that a unit receives is determined primarily by how unpredictable its $D_i$ value is, given its $X_i$ values. (See Appendix C for the mathematical details. The logic extends to maximum likelihood estimates of any generalized linear model.)

### Random Coefficient Models

Readers may wonder whether the reweighting associated with OLS is somehow ameliorated by random coefficient models, which in principle account for coefficient heterogeneity. There are many varieties of random coefficient models, and so to make the analysis more precise, consider the foundational model of Swamy (1970), which serves as the basis of many more computationally sophisticated models (Hsiao and Pesaran 2008). Suppose a unit $i$ is a member of one of $G$ groups indexed by $g$, where we use $g[i]$ to denote unit $i$'s group. Then the Swamy model supposes that

$$Y_i = \alpha_{g[i]} + \beta_{g[i]} D_i + X_i\gamma_{g[i]} + \epsilon_i. \qquad (10)$$

The objective is typically to estimate an average over groups (rather than units) of the group-specific coefficients; let us refer to this group-level average as $\bar{\beta}$. The standard estimator for the average of the group-specific coefficients is derived as the solution to a generalized least squares problem. Under the standard working assumption that the $\beta_g$s are independent draws from a distribution centered on $\bar{\beta}$, the estimator is a precision-weighted average of the group-specific OLS estimates (Hsiao and Pesaran 2008). Each of the group-specific OLS estimates is simply the value within each group of the quantity given by Expression (7). In effect, the estimator computes a series of quantities along the lines of Expression (7) and aggregates them on the basis of their precision. The precision weighting does not undo the sample distortions that the basic multiple regression weights introduce within each group. The effective sample is created from a combination of the precision weights and the within-group multiple regression weights.[17]

## Examples of Effective Samples

We can illustrate how these results apply to actual studies. We consider two examples. The first is based on Jensen (2003), who studies the effects of political regime type on inflows of foreign direct investment (FDI). This example shows how one may use our results to analyze a cross-national, time-series cross-section study. The second is based on Gerber and Huber (2010), who study the effects of partisanship on economic assessments. This example shows how one may use our results to analyze a survey-based study. Our purpose with these examples is not to expose flaws in the studies. Both studies are very well executed and, as a result, well cited. Rather, to the extent that we take unmodeled effect heterogeneity seriously, the goal is to illustrate how the multiple regression weights allow us to understand more precisely the scope conditions for the results from these studies. We find that the studies differ quite markedly in the extent to which the nominal samples reflect the effective samples used to generate key findings.

### Effects of Regime Type on FDI

Jensen (2003) studies the effects of regime type on FDI using cross-section, time-series cross-sectional, and other regression analyses of a set of 114 countries observed over the years 1970 to 1997. The results suggest that "democratic political institutions are associated with higher levels of FDI inflows" (588). The geographic distribution of the nominal sample is shown in the left panel of Figure 1. Our illustration focuses on the time-series cross-section

[17] When the assumption about the distribution of the $\beta_g$s is valid and either $D_i$ is randomly assigned within groups or $\tau_i$ is constant within groups, then the precision-weighted estimator is consistent for the average *over groups* of the group-level average causal effects, $\tau^G = \text{E}[\text{E}[\tau_i|g[i] = g]]$. This quantity is by construction not generally equal to the mean unit-level causal effect, $\bar{\tau}$.

**FIGURE 1 Example of nominal and effective samples from Jensen (2003)**


*Note:* On the left, the shading shows countries in the nominal sample for Jensen (2003) estimate of the effects of regime type on FDI. On the right, darker shading indicates that a country contributes more to the effective sample, based on the panel specification used in estimation.

analysis, and specifically on Model 10 in Table 4 of the published article. The specification incorporates country and decade fixed effects, lagged FDI, and a set of control variables including lagged values of market size, development level, growth, trade, budget deficit, government consumption, and democracy. The resulting estimate implies that a one-unit increase in the Polity score corresponds to a 0.020 increase in net FDI inflows as a percentage of gross domestic product ($p < 0.001$).

Of course, there is no way to randomly assign regime types to countries, and so estimates of the effects of regime types rely on the unexplained variation in regime type that the world provides us. That being said, not all countries contribute equally to such variation. The multiple regression weights measure such contributions. We first regress the FDI variable on the various control variables (including the country and decade fixed effects) and then extract the residuals. The multiple regression weights (the $\hat{w}_i$s) equal the squares of these residuals. To measure the extent to which a given country contributes to the effect estimate, we compute the analogue to the left-hand side of Expression (9), taking the sum of the weights for that country (summing over all time periods) and dividing it by the sum of all the weights in the sample. The right panel in Figure 1 shows the results. Darker shading means a larger contribution by that country, and white shading means a small or negligible contribution. From the sample of 114 countries, 12 contribute over half (51%) of the weight used to construct the estimate of the effect of regime type on FDI, and 32 contribute 90% of the weight. The top 12 contributing countries, in descending order of their weights, are Uruguay, Hungary, Niger, Philippines, Argentina, Madagascar, Pakistan, Zimbabwe, Poland, Peru, Lesotho, and Belarus. Importantly, the lowest-contributing 14 countries combine to contribute less than 0.05% of the weight used to construct the effect estimate. These lowest-contributing countries, in ascending order of their weights, are Central African Republic, Albania, Germany, Russia, Haiti, Benin, South Africa, Yemen, Honduras, DR Congo, Czech Republic, Lithuania, China, and Latvia.

These results are useful for two reasons. First, they provide a precise statement on the scope conditions for the findings. We see that the findings are driven primarily by the experiences of Latin American, Eastern European, and African cases, the exceptions being the Philippines and Pakistan. A substantively important finding is that none of the fast-growing economies of East Asia (the "Asian Tiger" economies of Hong Kong, Singapore, South Korea, and Taiwan, along with Indonesia, Malaysia, or Thailand) are major contributors to this estimate, even though they may be countries of interest for this study. In addition, large and theoretically interesting countries such as Russia, China, and Germany are essentially not represented at all in the estimate of the causal effect. Second, the results help in determining which cases one may want to investigate further to check the accuracy of one's interpretation of the quantitative results. This provides a refinement to the nested mixed-methods strategies proposed by Lieberman (2005) for using quantitative results to select cases for qualitative investigation.

## Effects of Partisanship on Economic Assessments

Gerber and Huber (2010) study the effects of partisanship on economic assessments by analyzing data from the

Cooperative Congressional Election Survey (CCES) collected just before and just after the 2006 midterm election. The research design took advantage of the unexpected Democratic takeover of the House and Senate as a result of the election. As such, the authors claim that correlations between partisanship and economic perceptions can be interpreted in terms of the differential effects of the partisan change in control of Congress on Republican- versus Democratic-leaning partisans. To rule out spuriousness with respect to other factors that affect economic perceptions, Gerber and Huber control for gender, race, age, union membership, income, and education. They find that Democratic partisans appear to be substantially more upbeat about economic conditions.

We focus on regression Model (4) in Table 4A of their article, which uses a behavioral indicator of economic assessments: the natural log of the change in respondents' predicted vacation spending over the pre- and post-election waves of the survey. The nominal sample is a subsample of 1,469 subjects from the overall CCES sample. The CCES sample included 2,000 respondents selected from a pool of Internet users matched to the characteristics of a national random-digit dialing sample. The subsample of 1,469 includes members of the panel for whom data were complete for this regression model. The characteristics of this nominal sample are shown in the second and third columns of Table 1.[18] Gerber and Huber(2010) estimate that a unit change in a respondent's party identification scale in the direction of "strong Democrat" is associated with a 0.077 increase in log change in projected vacation spending ($p < .05$), implying about a 20% increase.

To what extent does the nominal sample differ from the effective sample that generated this particular estimate? That question is answered in the fourth and fifth columns of Table 1, which show summary statistics for the effective sample as produced by regression. To compute these summary statistics, we regressed the party identification variable on all of the control variables, extracted the residuals, and took the squares of these residuals to obtain the $\hat{w}_i$s. The summary statistics for the effective sample are the weighted means (weighted by the $\hat{w}_i$s) for each of the variables, computed using the left-hand side of Expression (9). In this case, we see small differences between the nominal and effective samples. The reason is that party identification, the causal factor of interest, exhibits a high degree of unexplained variation relative to many of the demographic controls considered in this analysis. For controls that are more prognostic of party identification (e.g., the indicators for race and union membership), there is very little variation to start with in the nominal sample, and so reweighting on these controls is minor when we consider demographic shares in absolute terms. In this example, the nominal sample rather accurately characterizes the population for which the effect of partisanship has been estimated.[19]

## Estimating Representative Causal Effects

Researchers may want to do more than accept what regression offers in terms of multiple regression weighting and produce an estimate of $\bar{\tau}$, the average causal effect for the target population represented by the nominal sample. There is a large current literature on identification and estimation of average causal effects, and it is beyond the scope of this article to detail all of the results from this literature. We describe the basic conditions required to estimate representative causal effects, provide references that explain in more detail various approaches to operationalizing such conditions, and explain how multiple regression compares.

So far, we have used a causal model that is very favorable to regression, assuming both locally linear unit-level effects (Expression (1)) and linearity of $D_i$ with respect to the control variables (Expression (3)). Under these assumptions, we find that multiple regression produces a causal effect estimate that incorporates unit-level contributions on the basis of conditional variance of the treatment variable. One idea might be that we could just weight by the inverse of such conditional variances to recover $\bar{\tau}$. This would require obtaining consistent and stable estimates of the conditional variances. (We cannot weight by the inverse of the $\hat{w}_i$s because the unit-level weights may be arbitrarily close to zero, even as $n \to \infty$.) However, this inverse-variance weighting procedure is no less onerous than are the steps required to estimate average causal

[18]The summary statistics for the nominal sample that we show are slightly different from what Gerber and Huber (2010) show in Table 2 of their article, because we compute summary statistics only for the subsample with complete data and therefore that was actually used in the regression. This allows us to focus on differences between nominal and effective samples due solely to the multiple regression weights.

[19]Whether or not this makes Gerber and Huber's (2010) estimates "externally valid" depends on whether the nominal sample is representative of the target population for which one wants to make inferences. In this case, because of the manner in which the CCES sample was constructed and patterns of nonresponse in the CCES, there are significant differences in the demographic profile of the CCES sample and the population of voting-age Americans, for example. Such differences are distinct from the types of sample distortions that are of interest in the current discussion.

**TABLE 1 Summary Statistics for the Nominal and Effective Samples Relevant to Model (4) in Table 4A of Gerber and Huber Gerber and Huber (2010)**

| | Nominal Sample | | Effective Sample | |
|---|---|---|---|---|
| Variable | Mean | S.D. | Mean | S.D. |
| Party ID (Scale: −2 = Strong Republican to +2 = Strong Democrat) | 0.05 | 1.35 | −0.10 | 1.74 |
| Age (years) | 47.58 | 15.11 | 48.06 | 15.78 |
| Female (1 = yes) | 0.52 | 0.50 | 0.52 | 0.50 |
| Hispanic (1 = yes) | 0.04 | 0.21 | 0.06 | 0.23 |
| Black (1 = yes) | 0.04 | 0.19 | 0.03 | 0.16 |
| Union member (1 = yes) | 0.08 | 0.27 | 0.07 | 0.26 |
| Income (Scale: 0 to 1) | 0.59 | 0.26 | 0.58 | 0.25 |
| Income Refused/Don't Know | 0.11 | 0.31 | 0.09 | 0.29 |
| Education (Scale: 0 to 5) | 2.51 | 1.32 | 2.41 | 1.30 |
| Pre-election household income forecast | 0.46 | 0.96 | 0.54 | 0.94 |
| Post-election household income forecast | 0.42 | 0.98 | 0.48 | 0.95 |
| Pre-election national economy forecast | −0.11 | 0.97 | 0.04 | 1.01 |
| Post-election national economy forecast | −0.14 | 0.89 | −0.05 | 0.90 |
| Log change in holiday spending | −0.04 | 1.19 | −0.02 | 1.09 |
| Log change in vacation spending | 0.05 | 2.18 | 0.05 | 2.21 |
| Pre-election happiness | 1.93 | 0.83 | 2.01 | 0.84 |
| Post-election happiness | 1.95 | 0.76 | 2.02 | 0.75 |
| Pre-election state economy forecast | 0.04 | 0.84 | 0.11 | 0.85 |
| Post-election state economy forecast | −0.03 | 0.88 | 0.00 | 0.88 |

effects under assumptions much less restrictive than the two linearity assumptions.

Thus, let us consider identification and estimation of average causal effects under a more general model of treatment assignment and unit-level effects. The target quantity is again $\bar{\tau}$, as defined in Expression (4); however, we now drop the assumption of local linearity, as stated by Expression (1). We instead define causal effects over arbitrary potential outcomes:

$$\tau_i(d, d') = \frac{Y_i(d') - Y_i(d)}{d' - d},$$

where the $d$ and $d'$ in parentheses are to emphasize that causal effects are defined for treatment values $d$ and $d'$. These values would be chosen by the researcher on the basis of some meaningful comparison. Because we are allowing for effects to be potentially nonlinear, we need to pin down treatment values $d$ and $d'$ as the basis of estimation. (This is similar to what is commonly done to interpret nonlinear models, e.g., by considering the consequences of changing a regressor from its 25th to 75th percentile value.) Because local linearity may not hold, $\tau_i(d, d')$ may take on different values depending on the choice of $d$ and $d'$. Having pinned down our treatment values, we continue to assume that $X_i$ is sufficient to control for all sources confounding between $D_i$ and outcomes that would be realized under treatments $d$ or $d'$. This is expressed formally as

$$(Y_i(d), Y_i(d')) \perp\!\!\!\perp D_i | X_i. \tag{11}$$

Additionally, we no longer need to assume that the conditional expectation of $D_i$ is linear in the covariates. This general case is the subject of the literature on "generalized treatment regimes" (Imai and van Dyk 2004; Imbens 2000; Imbens and Wooldridge 2009).

Under this highly agnostic model of causal effects, estimation of $\mathrm{E}[\tau_i(d, d')]$ requires that a strong "positivity" condition holds over the whole population.[20] Positivity, loosely speaking, requires that, for all values of $X_i$ that appear in the target population, there is some probability of observing different values of $D_i$. If, for example, all units with a given covariate profile always have the same treatment condition, then one cannot estimate causal effects for these units. When positivity fails, then the best

[20]This assumption is typically coupled with the unconfoundedness assumption given by Expression (11)—for example, Imai and van Dyk (2004, Assumption 2).

that one can do without introducing more assumptions (that provide a basis for extrapolation and interpolation) is to estimate a representative causal effect for the subset of the target population for which positivity does hold (Petersen et al. 2011). Formally, the positivity assumption is as follows:

$$\Pr[D_i = d|X_i = x] > 0, \quad \Pr[D_i = d'|X_i = x] > 0$$

for all values of $x$ in the target population represented by the nominal sample.

Given unconfoundedness and positivity, the average causal effect for the target population is defined by

$$\bar{\tau}(d, d') = \mathrm{E}[\tau_i(d, d')] = \frac{\mathrm{E}[Y_i(d')] - \mathrm{E}[Y_i(d)]}{d' - d}.$$

Expectations ($\mathrm{E}[.]$) are taken with respect to the target population. To estimate $\bar{\tau}(d, d')$, we need only plug in estimates of $\mathrm{E}[Y_i(d)]$ and $\mathrm{E}[Y_i(d')]$. There are different ways to estimate such quantities. Some methods use estimates of $\Pr(D_i = d|X_i)$ and $\Pr(D_i = d'|X_i)$ to construct inverse probability weights. It is clear why positivity is important for such methods: When $\Pr(D_i = d|X_i)$ or $\Pr(D_i = d'|X_i)$ is zero, weights that use its inverse are undefined. Other methods flexibly model the relationship between $X_i$, $D_i$, and the potential outcomes to impute values of $Y_i(d)$ and $Y_i(d')$ for all units in the nominal sample.[21] A special case of such a method is a regression estimator where $D_i$ is interacted with $X_i$ in a flexible manner. Imbens and Rubin (2011, chap. 7) recommend such an interacted regression estimator wherein the $X_i$ variables have been centered about their means; this retargets the estimator so that the coefficient on $D_i$ directly estimates the population average causal effect. However, regardless of the method used, it is clear why we need positivity if we seek to estimate population average causal effects. If there are values of $X_i$ for which there are never any $Y_i(d)$ or $Y_i(d')$ values, then we have to rely on pure extrapolation or interpolation to estimate effects for those values of $X_i$. These different approaches are discussed in detail by Imbens and Wooldridge (2009) as well as in current textbooks on causal inference, such as Imbens and Rubin (2011), Hernan and Robins (2013), and Van der Laan and Rose (2011).

We can draw a connection between the positivity assumption and the results that we obtained above for multiple regression. The positivity assumption implies that $\mathrm{Var}[D_i|X_i = x] > 0$, for all values of $x$. Thus, when a multiple regression places (either approximately or exactly) no weight on a set of units with a given covariate profile, this result implies that positivity is violated and representative causal effects cannot be estimated for the target population without assumption-based extrapolation or interpolation. Examination of the effective sample provides insight about what can be learned from any given study. In the Jensen (2003) study, many substantively important countries—including Russia and China—cannot be included in estimates of the causal effects of regime type on FDI. No consistent, extrapolation-free estimator of $\mathrm{E}[\tau_i(d, d')]$ is available for any $(d, d')$ if our interest is in the average over the entirety of the 114 countries in the nominal sample. This is not due to any fault on Jensen's part; it is an implication of the variation that nature has provided. However, in other studies, such as Gerber and Huber (2010), positivity appears to hold, and thus estimation of representative average causal effects is straightforward. Of course, the effective sample that multiple regression generates is one for which positivity holds by construction. In conjunction with the linearity assumptions embodied by Expressions (1) and (3) (assumptions that are with little loss of generality when $D_i$ is binary), regression facilitates estimation in a manner that reflects the regions of the data where causal inference is possible.

[21] Van der Laan and Rose (2011, chap. 3) discuss machine learning algorithms for imputing potential outcomes.

## Conclusion

A key argument for using regression on representative samples to estimate causal effects rather than experiments or quasi-experimental methods using instrumental variables, discontinuities, or matching is based on the presumption that one can avoid problems of external validity. From the results above, we see that this argument is not generally valid even under circumstances that are highly favorable to regression. The mechanics of regression adjustment are such that calculation of effect estimates is based on sources of identifying variation in the data. The weighting that this implies may not map onto the population that the researcher sets out to study. Running a regression without considering the steps involved in the estimation obscures the nonrepresentative nature of the effects that have been estimated. Experimental and quasi-experimental methods are more transparent in that regard. This is to say nothing of the differences in the degree of internal validity that the different approaches achieve.

But our results are not nihilistic. To the extent that a regression has adequately controlled for confounding, we have demonstrated how to characterize the effective sample. In a world of heterogeneous effects, characterizing the effective sample is critical to the scientific enterprise of creating generalized knowledge about the effects of

causal factors. Such generalized knowledge requires that we understand how effects vary across different types of populations. We can only understand how effects vary across different types of populations if we know the profile of the samples that gave rise to our effect estimates. The multiple regression weights are precisely what one needs to characterize the profile of the effective sample that gave rise to an effect estimate from a regression study. The thousands of well-specified regressions that populate our journals are not without value. It is just that the scope of their applicability is probably more limited than what we have been led to believe in the absence of knowing the multiple regression weights. Researchers building on such past work would do well to investigate such scope conditions using the methods outlined in this article, creating maps and summary statistics that provide a clear picture of the effective sample. We may find that effects differ across effective samples that also differ in their covariate profiles. Such information on effects and associated effective sample characteristics provides the basis for building theories to explain effect heterogeneity and produce a generalized understanding of the effects of causal factors.

Readers may be disappointed that we have not proposed a statistical "fix" to generalizability problems for a given study. But it is crucial for researchers to understand what really is being estimated with the statistical methods that they use and that statistical methods themselves cannot overcome limitations of the data. We have shown that achieving knowledge about causal effects for a given population is more difficult than running a conventional multiple regression on a sample representative of that population. Estimates from a randomized experiment are only directly informative for the subpopulation whose treatment status can be manipulated by the investigator. Estimates from an observational study can only be directly informative for the subpopulation that exhibits some unpredictability in their treatment status after accounting for control variables.

Do these limits doom us to a world of idiosyncratic knowledge? The answer depends on the manner in which we conduct and interpret our empirical work. Abandoning experiments and quasi-experiments for regressions on representative samples is not a solution. As Imbens (2010) suggests, there are principled ways to move beyond merely accumulating idiosyncratic findings and analyzing effect heterogeneity across different studies. Furthermore, by understanding the effective sample for a study, one can inform theory. The effective sample for a study may give a lot of weight to units that existing theory suggests should not be affected by the treatment. If a substantial effect is nonetheless found, this creates an opportunity for refining current theories. It may be more fruitful to think about the *theoretical traction* that a study's sample offers rather than evaluating it against some illusive generalizability ideal. Regardless of the approach taken, an understanding of the effective sample is the first step in placing effect estimates into a broader context.

# Appendix A:
## Multiple Regression Weights Results

**Theorem A.1.** *Given the data-generating process outlined in the main text, namely, that the following hold:*

1. *(Unconfoundedness)* $(Y_i(0), \tau_i) \perp\!\!\!\perp D_i | X_i$.
2. *(Locally linear unit-level effects)* $Y_i = Y_{0i} + \tau_i D_i$.
3. *(Linearity of treatment in control vector)* $\mathrm{E}[D_i|X_i] = V_i\omega$, *with* $V_i = (1 \quad X_i)$ *and* $\omega$ *a* $K+1$ *column vector of coefficients.*
4. *The data are well behaved such that the OLS solutions converge to a limit.*

*Then as* $n \to \infty$, *the OLS estimate* $\hat{\beta}$ *for the coefficient* $\beta$ *in Expression* (6) *obeys*

$$\hat{\beta} \xrightarrow{p} \frac{\mathrm{E}[w_i\tau_i]}{\mathrm{E}[w_i]}, \qquad \text{where } w_i = (D_i - \mathrm{E}[D_i|X_i])^2.$$

*Proof.* Define $W_i = (1 \quad D_i \quad X_i)$, a $K+2$-length row vector combining the constant, causal factor, and control variables, and $\delta = (\alpha \quad \beta \quad \gamma)'$, a $K+2$-length column vector containing the model coefficients. Define the stacked matrix of regressors and outcome vector as

$$\mathbf{W} = \begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

The OLS estimates may be written as

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\gamma}_1 \\ \vdots \\ \hat{\gamma}_k \end{pmatrix} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'Y.$$

By partial regression,

$$\hat{\beta} = \frac{\sum_{i=1}^n \tilde{D}_i Y_i}{\sum_{i=1}^n \tilde{D}_i^2}, \tag{12}$$

where $\tilde{D}_i = D_i - V_i\hat{\omega}$, with $\mathbf{V} = (V_1' \quad \cdots \quad V_n')'$, $D = (D_1 \quad \cdots \quad D_n)'$, and $\hat{\omega} = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'D$. The partial regression solution is usually expressed in terms of residualized $Y_i$ and $D_i$ values, but that is equivalent to the expression given here because of the idempotence of the residual-maker matrix, $I - \mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'$ (Angrist and Pischke 2009, 36; Greene 2008, 29–31).

Substituting $(Y_{0i} + \tau_i D_i)$ for $Y_i$ in Expression (12), we have

$$\hat{\beta} = \frac{\sum_{i=1}^n \tilde{D}_i [Y_{0i} + \tau_i D_i]}{\sum_{i=1}^n \tilde{D}_i^2} = \frac{\sum_{i=1}^n \tilde{D}_i Y_{0i}}{\sum_{i=1}^n \tilde{D}_i^2} + \frac{\sum_{i=1}^n \tilde{D}_i D_i \tau_i}{\sum_{i=1}^n \tilde{D}_i^2} = \frac{\sum_{i=1}^n \tilde{D}_i Y_{0i}}{\sum_{i=1}^n \tilde{D}_i^2} + \frac{\sum_{i=1}^n \tilde{D}_i V_i \hat{\omega} \tau_i}{\sum_{i=1}^n \tilde{D}_i^2} + \frac{\sum_{i=1}^n \tilde{D}_i^2 \tau_i}{\sum_{i=1}^n \tilde{D}_i^2}. \quad (13)$$

By linearity of $D_i$ in $X_i$, $\hat{\omega} \xrightarrow{p} \omega$. To see that the first term on the right-hand side of Expression (13) goes to zero as $n \to \infty$, start by applying the weak law of large numbers and Slutsky's theorem:

$$\frac{\frac{1}{n}\sum_{i=1}^n \tilde{D}_i Y_{0i}}{\frac{1}{n}\sum_{i=1}^n \tilde{D}_i^2} = \frac{\frac{1}{n}\sum_{i=1}^n (D_i - V_i\hat{\omega}) Y_{0i}}{\frac{1}{n}\sum_{i=1}^n (D_i - V_i\hat{\omega})^2} \xrightarrow{p} \frac{\mathrm{E}[D_i Y_{0i}] - \mathrm{E}[(V_i\omega) Y_{0i}]}{\mathrm{E}[w_i]}.$$

Take the numerator, iterate expectations with respect to $X_i$, and apply unconfoundedness and linearity of $D_i$ in $X_i$:

$$\begin{aligned} \mathrm{E}[D_i Y_{0i}] - \mathrm{E}[(V_i\omega) Y_{0i}] &= \mathrm{E}_X[\mathrm{E}[D_i Y_{0i}|X_i]] \\ &\quad - \mathrm{E}_X[(V_i\omega)\mathrm{E}[Y_{0i}|X_i]] \\ &= \mathrm{E}_X[\mathrm{E}[D_i|X_i]\mathrm{E}[Y_{0i}|X_i]] \\ &\quad - \mathrm{E}_X[(V_i\omega)\mathrm{E}[Y_{0i}|X_i]] = 0. \end{aligned} \quad (15)$$

We now consider the second term in the right-hand side of Expression (13), where

$$\frac{\sum_{i=1}^n \tilde{D}_i V_i \hat{\omega} \tau_i}{\sum_{i=1}^n \tilde{D}_i^2} \xrightarrow{p} \frac{\mathrm{E}[(D_i - V_i\omega) V_i \omega \tau_i]}{\mathrm{E}[w_i]}.$$

Again considering the numerator,

$$\begin{aligned} \mathrm{E}[(D_i - V_i\omega) V_i \omega \tau_i] &= \mathrm{E}_X[V_i \omega \mathrm{E}[(D_i - V_i\omega)\tau_i|X_i]] \\ &= 0, \end{aligned}$$

by unconfoundedness and linearity of $D_i$ in $X_i$.

This leaves us to consider the limiting behavior of the third term in the right-hand side of Expression (13). What remains is a weighted average of the $\tau_i$s, which by the law of large numbers and Slutsky's theorem obeys

$$\frac{\sum_{i=1}^n \tilde{D}_i^2 \tau_i}{\sum_{i=1}^n \tilde{D}_i^2} = \frac{\sum_{i=1}^n (D_i - V_i\hat{\omega})^2 \tau_i}{\sum_{i=1}^n (D_i - V_i\hat{\omega})^2} \xrightarrow{p} \frac{\mathrm{E}[w_i\tau_i]}{\mathrm{E}[w_i]}. \quad \square$$

**Corollary A.1.** *Given the data-generating process outlined in the main text, if $D_i$ is randomly assigned, then as $n \to \infty$,*

$$\hat{\beta} \xrightarrow{p} \bar{\tau}.$$

*Proof.* Under random assignment, $w_i \perp \tau_i$. Then, by random sampling and Slutsky's theorem,

$$\frac{\sum_{i=1}^n \tilde{D}_i^2 \tau_i}{\sum_{i=1}^n \tilde{D}_i^2} \xrightarrow{p} \frac{\mathrm{E}[w_i]\mathrm{E}[\tau_i]}{\mathrm{E}[w_i]} = \mathrm{E}[\tau_i] = \bar{\tau}. \quad \square$$

# Appendix B:

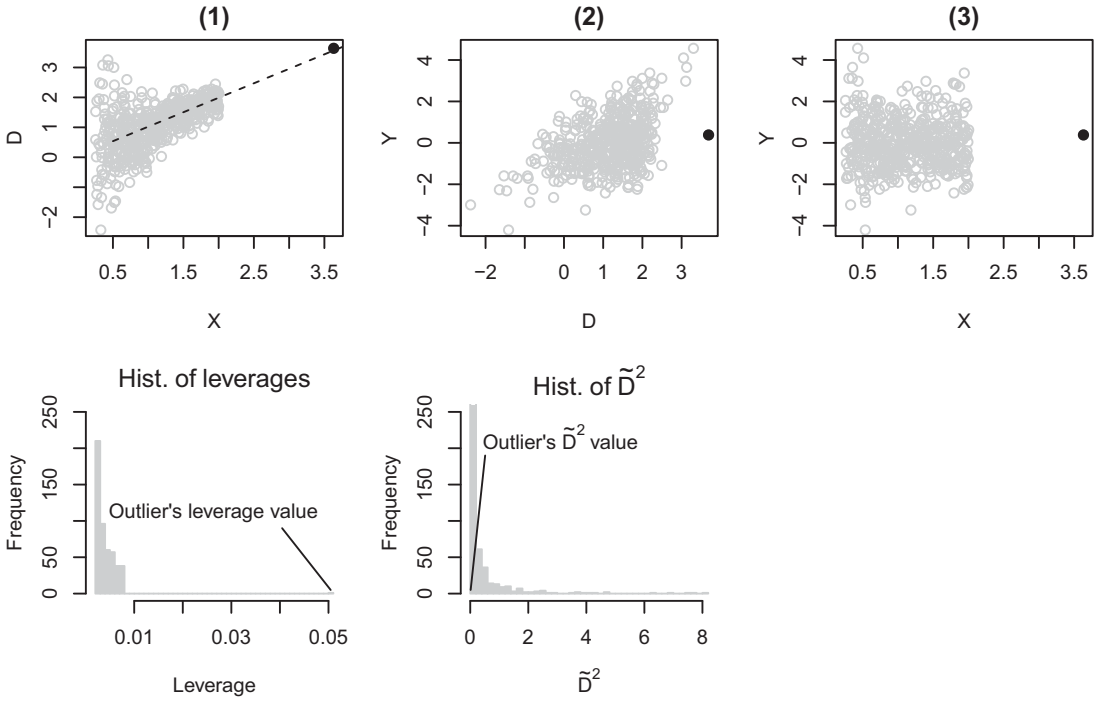
## Leverage versus Multiple Regression Weights

The multiple regression weights isolate units' contributions to the estimate of the causal effect of the treatment, $D$, whereas the leverage measures influence associated with all regressors, including both $D$ and the covariates. As such, the multiple regression weights and the leverage can diverge in the amount of weight they assign to a unit. Figure B1 uses a simulation to provide a clear illustration. In this case, the correct control vector includes the scalar random variable $X$. The distribution of $X$ is such that there is one outlier. This outlier is colored black in the top row of graphs. Given how far this data point is from the rest of data on the $X$ dimension (and, by implication, the $D$ dimension), it is an extremely high leverage data point. This is shown in the histogram in the bottom left of Figure B1. Its leverage statistic value is extreme relative to other data points. However, Scatterplot (1) shows that this outlier resides very close to the regression line for the regression of $D$ on $X$. As such, its estimated multiple regression weight—$\tilde{D}^2$, being the square of the residual from the regression of $D$ on $X$—is very small, as shown in the histogram on the right. The units with large multiple regression weights are those far from the regression line in Scatterplot (1)—specifically, those with $X$ values close to 0.5.

Another example may also help to clarify how relying on the leverage can have one draw precisely the opposite conclusion that one should draw about what types of units (defined by a binary covariate) are contributing the most to a causal effect estimate. Suppose we had a binary treatment, $D_i$, and a single binary covariate, $X_i$, with the joint distribution characterized by

$$\begin{aligned} \Pr[D_i = 0, X_i = 0] &= 0.05, \\ \Pr[D_i = 1, X_i = 0] &= 0.20, \\ \Pr[D_i = 0, X_i = 1] &= 0.55, \\ \Pr[D_i = 1, X_i = 1] &= 0.20. \end{aligned}$$

**FIGURE B1 Multiple regression weights versus the leverage**




*Note:* In the top row, panel (1) shows a scatter plot of $D$ on $X$ with a regression fit, panel (2) shows the scatter plot of $Y$ on $D$, and panel (3) shows the scatter plot of $Y$ on $X$. The black dot is an outlier in the dimension of $X$. In the bottom row, the left panel is a histogram of the leverages from the regression of $Y$ on $D$ and $X$. The right panel shows the squared residuals from the regression of $D$ on $X$. In both histograms, the value for the outlier is flagged.

Applying Expression (8), the multiple regression weights obey

$$\mathrm{E}[w_i|X_i = 0] = \mathrm{Var}[D_i|X_i = 0] = 0.160,$$
$$\mathrm{E}[w_i|X_i = 1] = \mathrm{Var}[D_i|X_i = 1] = 0.196,$$

so that the causal effects for units with $X_i = 1$ receive 22% more weight in the construction of $\hat{\beta}$ than do those for units with $X_i = 0$.

Now we consider the leverage. The leverage for unit $i$ is

$$h_{ii} = W_i(\mathbf{W}'\mathbf{W})^{-1}W_i',$$

where $W_i$ and $\mathbf{W}$ are defined as in Appendix A. After normalizing by $n$, straightforward mathematical calculations obtain

$$\mathrm{E}[n \cdot h_{ii}|D_i = 0,\ X_i = 0] = 7.43,$$
$$\mathrm{E}[n \cdot h_{ii}|D_i = 1,\ X_i = 0] = 4.21,$$
$$\mathrm{E}[n \cdot h_{ii}|D_i = 0,\ X_i = 1] = 1.71,$$
$$\mathrm{E}[n \cdot h_{ii}|D_i = 1,\ X_i = 1] = 4.21.$$

Applying the law of iterated expectations,

$$\mathrm{E}[n \cdot h_{ii}|X_i = 0] = 4.86,$$
$$\mathrm{E}[n \cdot h_{ii}|X_i = 1] = 2.38,$$

so that units with $X_i = 1$ have, on average, 51% *less* leverage than units with $X_i = 0$ have. Thus, leverage-based calculations can be misleading if researchers are interested in the weighting of causal effects associated with $\hat{\beta}$.

# Appendix C:

## Maximum Likelihood Estimation for Nonlinear Models

We illustrate the case of implicit weighting for a logistic regression with treatment, $D_i$, and a vector of control variables, $X_i$. The illustration relies on a first-order approximation of the inverse link function for logistic regression. The approximation is necessary because there is no closed-form solution to the coefficient vector from a logistic regression. While our illustration is for logistic regression, the results hold for any nonlinear model that can be expressed in terms of estimating equations in the manner that follows (Liang and Zeger 1986). This holds for generalized linear models such as probit, Poisson regression, ordered logit or probit, multinomial logit or probit, and duration models (McCullagh and Nelder 1999). To apply the result below for these other models,

one substitutes the relevant predicted mean or inverse link function into the estimating equation.

Consider the following logistic regression model:

$$\Pr[Y_i = 1|D_i, X_i] = \text{logit}^{-1}(\alpha_\ell + \beta_\ell D_i + X_i\gamma_\ell),$$

where $\text{logit}^{-1}(a) = [1 + \exp(-a)]^{-1}$. We fit the model via maximum likelihood (ML) to obtain the coefficient vector, $\hat{\theta} = (\hat{\alpha}_\ell \quad \hat{\beta}_\ell \quad \hat{\gamma}_\ell)'$. As before, let $W_i = (1 \quad D_i \quad X_i)$. The ML estimator for $\hat{\theta}$ solves the following estimating equation (Liang and Zeger 1986):

$$\begin{aligned}\sum_{i=1}^{n} \psi(Y_i, W_i, \hat{\theta}) &= \sum_{i=1}^{n} \left(W_i\{Y_i - \text{logit}^{-1}(W_i\hat{\theta})\}\right) \\ &= \mathbf{0}_{1\times(K+2)},\end{aligned}$$

where $K$ is the number of variables in $X_i$.

We may approximate the $\text{logit}^{-1}(.)$ function by taking a first-order Taylor expansion about $\text{logit}(P_Y) = \log[P_Y/(1 - P_Y)]$, where $P_Y = \Pr[Y_i = 1]$:

$$\begin{aligned}\text{logit}^{-1}(a) &\approx P_Y + \frac{\exp[\text{logit}(P_Y)]}{[1 + \text{logit}(P_Y)]^2}[a - \text{logit}(P_y)] \\ &= b + ca,\end{aligned}$$

where $b = P_Y - P_Y(1 - P_Y)\text{logit}(P_Y)$ and the scale factor $c = P_Y(1 - P_Y)$.

Using the linearized $\text{logit}^{-1}(.)$ function, we may define approximate estimating equations:

$$\begin{aligned}\sum_{i=1}^{n} \psi(Y_i, W_i, \hat{\theta}) &\approx \sum_{i=1}^{n} \tilde{\psi}(Y_i, W_i, \hat{\theta}) \\ &= \sum_{i=1}^{n} \left[W_i(Y_i - b - cW_i\hat{\theta})\right] \\ &= \mathbf{0}_{1\times(K+2)}.\end{aligned}$$

Without loss of generality, we may reframe the problem by defining an alternative coefficient vector $\tilde{\theta} = ((b + c\hat{\alpha}_\ell) \quad c\hat{\beta}_\ell \quad c\hat{\gamma}_\ell)'$, so that

$$\sum_{i=1}^{n} \tilde{\psi}(Y_i, W_i, \hat{\theta}) = \sum_{i=1}^{n} \left[W_i(Y_i - W_i\tilde{\theta})\right] = \mathbf{0}_{1\times(K+2)},$$

which are precisely the estimating equations for the OLS estimator. Thus, the only difference between the OLS estimate and the approximate logistic regression estimate of $\beta_\ell$ is the scale factor $c$. Therefore, the reweighting results for OLS hold—as a first-order approximation—for the logistic regression maximum likelihood estimator of $\beta_\ell$. As mentioned above, analogous results can be derived for the maximum likelihood estimators of the coefficients in other generalized linear models.

# References


Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge. 2010. "Finite Population Causal Standard Errors." Unpublished manuscript, Harvard University, Stanford University, and Michigan State University.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics* (Vol. 3), ed. Orley C. Ahsenfelter and David Card. Amsterdam: North Holland, 1277–366.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Arceneaux, Kevin, and David W. Nickerson. 2009. "Who Is Mobilized to Vote? A Re-Analysis of Seven Randomized Field Experiments." *American Journal of Political Science* 53(1): 1–16.

Aronow, Peter M., and Joel A. Middleton. 2013. "A Class of Unbiased Estimators of the Average Treatment Effect in Randomized Experiments." *Journal of Causal Inference* 1(1): 135–54.

Banerjee, Abhijit V., and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1: 151–78.

Bardhan, Pranab. 2013. "Little, Big: Two Ideas about Fighting Global Poverty." *Boston Review*, May/June. http://www.bostonreview.net/world-books-ideas/pranab-bardhan-little-big

Barron, Reuben M., and David A. Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51(6): 1173–82.

Beck, Nathaniel, and Jonathan N. Katz. 2001. "Throwing Out the Baby with the Bath Water: A Comment on Green, Kim, and Yoon." *International Organization* 55(2): 487–95.

Berk, Richard A., Bruce Western, and Robert E. Weiss. 1995. "Statistical Inference for Apparent Populations." *Sociological Methodology* 25: 421–58.

Breusch, Trevor S., and Adrian R. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* 47(5): 1287–94.

Campbell, Donald T. 1957. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychology Bulletin* 45(4): 297–312.

Campbell, Donald T. 1984. "Can We Be Scientific in Applied Social Science?" *Evaluation Studies Review Annual* 9: 26–48.

Chamberlain, Gary. 1984. "Panel Data." In *Handbook of Econometrics* (Vol. 2), ed. Zvi Griliches and Michael D. Intriligator. New York: Elsevier, 1247–318.

Colgan, Jeff D. 2010. "Oil and Revolutionary Governments: Fuel for International Conflict." *International Organization* 64(4): 661–94.

Davidson, Russell, and James G. MacKinnon. 2004. *Econometric Theory and Methods*. Oxford: Oxford University Press.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48: 424–55.

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61(2): 282–93.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design and Analysis*. New York: Norton.

Gerber, Alan S., and Gregory A. Huber. 2010. "Partisanship, Political Control, and Economic Assessments." *American Journal of Political Science* 54(1): 153–73.

Gill, Jeff. 2001. "Whose Variance Is It Anyway? Interpreting Empirical Models with State-Level Data." *State Politics and Policy Quarterly* 1(3): 318–38.

Greene, William H. 2008. *Econometric Analysis*. 6th ed. Upper Saddle River, NJ: Pearson.

Heckman, James J., and Sergio Urzua. 2010. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." *Journal of Econometrics* 156(1): 27–37.

Hernan, Miguel A., and James M. Robins. 2013. *Causal Inference*. Boca Raton, FL: Chapman and Hall/CRC.

Hidalgo, F. Daniel, Suresh Naidu, Simeon Nichter, and Neal Richardson. 2010. "Economic Determinants of Land Invasions." *Review of Economics and Statistics* 92(3): 505–23.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3): 199–236.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945–60.

Hsiao, Cheng, and M. Hashem Pesaran. 2008. "Random Coefficient Models." In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice* (3rd ed.), ed. Laszlo Matyas and Patrick Sevestre. New York: Springer, 185–213.

Humphreys, Macartan. 2009. "Bounds on Least Squares Estimates of Causal Effects in the Presence of Heterogenous Assignment Probabilities." Unpublished manuscript, Columbia University.

Imai, Kosuke, and In Song Kim. 2012. "Understanding and Improving Linear Fixed Effects Regression Models for Causal Inference." Unpublished manuscript, Princeton University.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. "Misunderstandings between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society*, Series A 171(2): 481–502.

Imai, Kosuke, and David A. van Dyk. 2004. "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99(467): 854–66.

Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87: 706–10.

Imbens, Guido W. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48: 399–423.

Imbens, Guido W., and Donald B. Rubin. 2011. "Causal Inference in Statistics and Social Sciences." Unpublished manuscript, Harvard University.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47(1): 5–86.

Jensen, Nathan M. 2003. "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment." *International Organization* 57: 587–616.

King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2): 131–59.

Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73(1): 13–22.

Lieberman, Evan S. 2005. "Nested Analysis as a Mixed-Method Strategy for Comparative Research." *American Political Science Review* 99(3): 435–52.

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

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

McCullagh, Peter, and John A. Nelder. 1999. *Generalized Linear Models*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to Lab*. Cambridge: Cambridge University Press.

Neyman, Jerzy Splawa. [1923] 1990. "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9 (reprint)." *Statistical Science* 5(4): 465–72.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press.

Petersen, Maya L., Kristin E. Porter, Susan Gruber, Yue Wang, and Mark J. Van der Laan. 2011. "Positivity." *In Targeted Learning: Causal Inference for Observational and Experimental Data*, ed. Mark J. Van der Laan and Sherri Rose. New York: Springer, 162–86.

Robinson, Gregory, John E. McNulty, and Jonathan S. Krasno. 2009. "Observing the Counterfactual? The Search for Political Experiments in Nature." *Political Analysis* 17(4): 341–57.

Rosenbaum, Paul R. 1984. "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society*, Series A 147(5): 656–66.

Rosenbaum, Paul R. 1999. "Choice as an Alternative to Control in Observational Studies." *Statistical Science* 14(3): 259–304.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6(1): 34–58.

Rubin, Donald B. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *Journal of the American Statistical Association* 75(371): 591–93.

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer.

Swamy, P. A. V. B. 1970. "Efficient Inference in a Random Coefficient Regression Model." *Econometrica* 38(2): 311–23.

Van der Laan, Mark, and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Obervational and Experimental Data*. New York: Springer.

VanderWeele, Tyler J., and James M. Robins. 2007. "Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs." *Epidemiology* 18(5): 561–68.

Wooldridge, Jeffrey M. 2009. *Introductory Econometrics: A Modern Approach*. 4th ed. Mason, OH: South-Western.