What Are We Weighting For?

Gary Solon
Steven J. Haider
Jeffrey M. Wooldridge

ABSTRACT

When estimating population descriptive statistics, weighting is called for if needed to make the analysis sample representative of the target population. With regard to research directed instead at estimating causal effects, we discuss three distinct weighting motives: (1) to achieve precise estimates by correcting for heteroskedasticity; (2) to achieve consistent estimates by correcting for endogenous sampling; and (3) to identify average partial effects in the presence of unmodeled heterogeneity of effects. In each case, we find that the motive sometimes does not apply in situations where practitioners often assume it does.

I. Introduction

At the beginning of their textbook's section on weighted estimation of regression models, Angrist and Pischke (2009, p. 91) acknowledge, "Few things are as confusing to applied researchers as the role of sample weights. Even now, 20 years post-Ph.D., we read the section of the Stata manual on weighting with some dismay." After years of discussing weighting issues with fellow economic researchers, we know that Angrist and Pischke are in excellent company. In published research, top-notch empirical scholars make conflicting choices about whether and how to weight and often provide little or no rationale for their choices. And in private discussions, we have found that accomplished researchers sometimes own up to confusion or declare demonstrably faulty reasons for their weighting choices.

Our purpose in writing this paper is to dispel confusion and dismay by clarifying the issues surrounding weighting. Our central theme is that the confusion stems from

Gary Solon, Steven J. Haider, and Jeffrey M. Wooldridge are professors of economics at Michigan State University. *They are grateful for helpful comments from the referees, David Autor, Raj Chetty, John DiNardo, Todd Elder, Mike Elsby, Osborne Jackson, Fabian Lange, Jason Lindo, Jim Ziliak, and seminar participants at Michigan State University, Abt Associates, Drexel University, the University of Kentucky, and Tulane University. The data used in this article can be obtained beginning November 2015 through October 2018 from Gary Solon, Department of Economics, Michigan State University, East Lansing, MI 48824, solon@msu.edu.*
[Submitted February 2013; accepted September 2013]
ISSN 0022-166X E-ISSN 1548-8004 © 2015 by the Board of Regents of the University of Wisconsin System

a lack of clarity about which among multiple potential motives for weighting pertains to the research project at hand. Once one specifies the particular motive for weighting, it becomes straightforward to consider whether the purpose for weighting really does apply, to use appropriate diagnostics to check whether it does, and then to proceed with appropriate estimation and inference methods. Hence the title of our paper: “What Are We Weighting For?”

In the next section, we pose a prior question: “What Are We Trying to Estimate?” In some projects, the purpose is to estimate descriptive statistics for a particular population. In those cases, whether weighting is called for depends simply on whether weighting is necessary to make the analysis sample representative of the target population. But in many other projects, the purpose is to estimate causal effects. In those cases, the weighting issue becomes more nuanced.

In Sections III, IV, and V, we successively discuss three distinct potential motives for weighting when estimating causal effects: (1) to achieve more precise estimates by correcting for heteroskedasticity; (2) to achieve consistent estimates by correcting for endogenous sampling; and (3) to identify average partial effects in the presence of heterogeneous effects.[1] In each case, after explaining the potential relevance of the motive, we will note that the motive sometimes does not apply in situations where practitioners often assume it does. We will recommend diagnostics for assessing the advisability of weighting, and we will suggest methods for appropriate inference. In Section VI, we will summarize our analysis and our recommendations for empirical practice.

II. What Are We Trying to Estimate?

A. Descriptive Statistics for a Population

Sometimes the purpose of a research project is to estimate descriptive statistics of interest for a population. Consider, for example, the 1967 poverty rate for the United States, which was officially measured as 13 percent based on the Current Population Survey (U.S. Bureau of the Census 1968). But suppose that one sought to estimate that rate on the basis of the reports of 1967 income in the first wave of the Panel Study of Income Dynamics (PSID) in 1968. The complication is that the PSID began with a sample that purposefully overrepresented low-income households by incorporating a supplementary sample drawn from households that reported low income to the Survey of Economic Opportunity in 1967. As in other surveys that purposefully sample with different probabilities from different parts of the population, the point of the oversampling was to obtain more precise information on a subpopulation of particular interest—in this case, the low-income population.[2]

[1] We do not mean to imply that this list is exhaustive. For example, we will give little attention here to the use of propensity-score weighting to control for covariates when estimating treatment effects. That topic is discussed in Chapter 21 of Wooldridge (2010).

[2] Similarly, the Current Population Survey oversamples in less-populous states, and the first wave of the Health and Retirement Study oversampled blacks, Mexican-Americans, and residents of Florida.

If one estimated the 1967 poverty rate for the United States population with the poverty rate for the full PSID sample, without any weighting to adjust for the low-income oversample, one would estimate the U.S. poverty rate at 26 percent.[3] That, of course, is an upward-biased estimate because the PSID, by design, overrepresents the poor. But one might achieve unbiased and consistent estimation by using the PSID sample's weighted poverty rate, weighting by the inverse probabilities of selection.[4] A visualization of how this works is that the PSID sample design views the U.S. population through a funhouse mirror that exaggerates the low-income population. Weighted estimation views the sample through a reverse funhouse mirror that undoes the original exaggeration. It turns out that the PSID's weighted poverty rate is 12 percent, a more reasonable estimate than the 26 percent figure.

The poverty-rate example illustrates the simple case of estimating a population mean on the basis of a sample that systematically fails to represent the target population but can be made to represent it by weighting. Much economic research, however, seeks to estimate more complex population statistics. Suppose, for example, that one wishes to characterize the 1967 earnings gap between black and white men with the same years of schooling and potential work experience (age minus years of schooling minus five). Many labor economists would describe that gap by attempting to estimate the population linear projection of log earnings on a dummy variable that equals one for blacks along with controls for years of schooling and a quartic in potential experience.[5]

Now suppose that one estimates that population regression by performing ordinary least squares (OLS) estimation of the regression of log earnings on the race dummy, years of schooling, and a quartic in potential earnings for black and white male household heads in the PSID sample. Doing so estimates the coefficient of the dummy variable for blacks at –0.344. Because exp (–0.344) = 0.71, this estimate seems to imply that, among male household heads with the same education and potential experience, blacks tended to earn only 71 percent as much as whites.

As in the example of estimating the poverty rate, however, this estimate of the population linear projection might be distorted by the PSID's oversampling of low-income households, which surely must lead to an unrepresentative sample with respect to male household heads' earnings. But again, one can apply a reverse funhouse mirror by using weights. In particular, instead of applying ordinary (that is, equally weighted) least squares to the sample regression, one can use weighted least squares (WLS), minimizing the sum of squared residuals weighted by the inverse probabilities of selection. Doing so leads to an estimated coefficient of –0.260 for the dummy variable for blacks, implying that, among male household heads with the same education and potential experience, blacks tended to earn 77 percent as much as whites. This is still a

[3] This calculation is based on approximating the official poverty line by dividing the PSID-reported "needs standard" by 1.25.

[4] For simplicity, we are overlooking complications from nonresponse, including mishaps in the PSID's implementation of the low-income oversample. For discussion of the latter and further references, see Shin and Solon (2011, Footnote 11)

[5] This is not the only descriptive approach available. One of the alternatives would be to control for race differences in covariates through propensity-score weighting. See Chapter 21 of Wooldridge (2010) for a general discussion, and Elder, Goddeeris, and Haider (2011) for an application to the black/white difference in infant mortality rates.

large shortfall but not as large as implied by the OLS estimate. A likely reason is that the particular way that the PSID overrepresented the low-income population involved an especially concentrated oversampling of low-income households in nonmetropolitan areas of the South. The unweighted PSID therefore may understate typical income for blacks even more than for whites.

What both our examples have in common is that they involve estimating descriptive statistics for a population on the basis of sample data. If the sample is representative of the target population (the most straightforward case being a simple random sample drawn from that population), the population statistic is consistently estimated by the analogous sample statistic. If the sample is systematically unrepresentative of the population, the population statistic generally is not consistently estimated by the analogous sample statistic. But, if the way in which the sample is unrepresentative is known, the population statistic can be consistently estimated by reweighting the sample statistic with the inverse probabilities of selection.[6]

This point is intuitive and not at all controversial. So why does the issue of weighting provoke confusion and dismay among economic researchers? The answer, which will occupy the rest of this paper, is that much economic research is directed not at estimating population descriptive statistics but at estimating causal effects.

B. Causal Effects

In the microeconometrics textbooks of both Wooldridge (2010) and Angrist and Pischke (2009), the very first page describes the estimation of causal effects as the principal goal of empirical microeconomists. According to Angrist and Pischke, "In the beginning, we should ask, *What is the causal relationship of interest?* Although purely descriptive research has an important role to play, we believe that the most interesting research in social science is about questions of cause and effect, such as the effect of class size on children's test scores. . . ." Similarly, the first sentences in the Wooldridge textbook are, "The goal of most empirical studies in economics and other social sciences is to determine whether a change in one variable, say *w*, causes a change in another variable, say *y*. For example, does having another year of education cause an increase in monthly salary? Does reducing class size cause an improvement in student performance? Does lowering the business property tax rate cause an increase in city economic activity?"

In contrast to the case of estimating population descriptive statistics, when economists perform estimation of causal effects, the question of whether to weight the data is complex. There are several distinct reasons that we may (or, as we will stress, sometimes may not) prefer to use weights in our estimation. We will take up three distinct reasons respectively in the next three sections.

III. Correcting for Heteroskedasticity

One motivation for weighting, taught for decades in undergraduate and graduate econometrics classes, is to correct for heteroskedastic error terms and thereby achieve more precise estimation of coefficients in linear or nonlinear regres-

[6] For a general formal demonstration, see Wooldridge (1999).

sion models of causal effects. A nice example of this motivation comes from the literature on the impact of unilateral divorce laws on the divorce rate. During the 1970s, many U.S. states adopted laws allowing unilateral divorce instead of requiring mutual consent of both spouses. Were these laws responsible for the rise in divorce rates that occurred during that period? In two insightful and influential articles published in the *American Economic Review*, Leora Friedberg (1998) and Justin Wolfers (2006) reported differences-in-differences estimates of the impact of unilateral divorce laws on divorce rates. In particular, using a panel of annual state divorce rates over time, they estimated linear regressions of the divorce rate on dummy variables for unilateral divorce laws with controls for state fixed effects and secular time trends. Following the practice of many other top-notch empirical economists,[7] both Friedberg and Wolfers weighted by state/year population in the estimation of their regression models. Friedberg justified the weighting as a correction for population-size-related heteroskedasticity in the state/year error terms.

Table 1 presents examples from a wide set of variations on the Friedberg/Wolfers regressions reported in Lee and Solon (2011). The regressions are estimated with Wolfers' 1956–88 data on annual divorce rates by state. The main point of Wolfers' article was that the short-run and long-run effects of unilateral divorce may differ, so the regressions in Table 1 follow Wolfers in representing unilateral divorce with a set of dummy variables for whether unilateral divorce had been in place for up to two years, three to four years, five to six years, . . . , 13–14 years, or at least 15 years. The dependent variable is the logarithm of the annual divorce rate by state, and the regressions include controls for state fixed effects, year fixed effects, and state-specific linear time trends.

The table's first column follows Friedberg and Wolfers in estimating by weighted least squares with weighting by state/year population. The second column uses ordinary least squares, which weights all observations equally. In both instances, to maintain agnosticism about which weighting approach—if either—comes close to producing a homoskedastic error term, Table 1 reports standard errors robust to heteroskedasticity (as well as to serial correlation over time within the same state).[8]

Setting aside other interesting aspects of these results (for example, the absence in this specification of any evidence for a positive effect of unilateral divorce on divorce rates), notice this striking pattern: Even though Friedberg's expressed purpose in weighting was to improve the precision of estimation, the robust standard errors are smaller for OLS than for WLS. For the estimated effects over the first eight years after adoption of unilateral divorce, the robust standard errors for OLS are only about half those for WLS. Apparently, weighting by population made the estimates *much less* precise!

As discussed in Dickens (1990), harming the precision of estimation with this sort of weighting is quite a common phenomenon. To understand why, start with the classic heteroskedasticity-based argument for weighting when the dependent variable is

[7] Some other prominent examples of similarly weighted estimation are Card and Krueger (1992); Autor, Katz, and Krueger (1998); Levitt (1998); Donohue and Levitt (2001); Borjas (2003); and Dehejia and Lleras-Muney (2004).

[8] Lee and Solon (2011) show that, for both the OLS and WLS results, naïve standard errors that correct for neither heteroskedasticity nor serial correlation are far smaller than the robust ones. This occurs mainly because the error term is highly serially correlated.

Table 1
Estimated Effects of Unilateral Divorce Laws

	1	2
Dependent Variable:	Log of Divorce Rate	Log of Divorce Rate
Estimation Method:	WLS	OLS
First 2 years	–0.022	–0.017
	(0.063)	(0.026)
Years 3–4	–0.049	–0.014
	(0.063)	(0.031)
Years 5–6	–0.051	–0.022
	(0.064)	(0.034)
Years 7–8	–0.033	–0.013
	(0.065)	(0.039)
Years 9–10	–0.052	–0.030
	(0.067)	(0.046)
Years 11–12	–0.051	–0.015
	(0.074)	(0.052)
Years 13–14	–0.043	–0.005
	(0.077)	(0.060)
Years 15+	0.006	0.026
	(0.084)	(0.073)

Notes: These results are drawn from Lee and Solon (2011, Table 2). The divorce rate is the number of divorces per 1,000 persons by state and year. The standard errors in parentheses are robust to heteroskedasticity and serial correlation. Both regressions include controls for state fixed effects, year fixed effects, and state-specific linear time trends.

a group average and the averages for different groups are based on widely varying within-group sample sizes. Simplifying by focusing on a cross-sectional example, suppose the model to be estimated is

$$(1)\quad y_i = X_i\beta + v_i$$

where y_i is a group-level average outcome observed for group i and the error term is fully independent of the explanatory variables. The group-average error term v_i equals $\Sigma_{j=1}^{J_i} v_{ij} / J_i$, where v_{ij} is the microlevel error term for individual j in group i and J_i denotes the number of individuals observed in group i. If v_{ij} is independently and identically distributed with variance σ^2, then elementary statistics shows that the variance of the group-average error term v_i is σ^2 / J_i. Thus, if J_i varies widely across groups (for example, if many more individuals are observed in California than in Wyoming), the group-average error term v_i is highly heteroskedastic. Then, as taught in almost every introductory econometrics course, OLS estimation of β in Equation 1 is inefficient and also leads to inconsistent standard errors if nothing is done to correct the standard errors for heteroskedasticity. The WLS estimator that applies least squares to the reweighted equation

$$(2) \quad \sqrt{J_i}y_i = \sqrt{J_i}X_i\beta + \sqrt{J_i}v_i$$

is the minimum-variance linear unbiased estimator and also generates consistent standard errors.

This presumably is the line of thinking that led Friedberg and Wolfers to use WLS to estimate their divorce-rate regressions. Compared to Wyoming, California offers many more observations of the individual-level decision of whether or not to divorce, and therefore it seems at first that weighting by state population should lead to more precise coefficient estimation. And yet, for the specification shown in Table 1, it appears that weighting by population *harms* the precision of estimation.

Dickens' (1990) explanation of what is going on is that, in many practical applications, the assumption that the individual-level error terms v_{ij} are independent is wrong. Instead, the individual-level error terms within a group are positively correlated with each other because they have unobserved group-level factors in common. In current parlance, the individual-level error terms are "clustered." Dickens illustrates with the simple example of an error components model for the individual-level error term:

$$(3) \quad v_{ij} = c_i + u_{ij}$$

where each of the error components, c_i and u_{ij}, is independently and identically distributed (including independence of each other), with respective variances σ_c^2 and σ_u^2. In this scenario, the variance of the group-average error term v_i is not σ^2 / J_i, but rather is

$$(4) \quad Var(v_i) = \sigma_c^2 + (\sigma_u^2 / J_i).$$

If σ_c^2 is substantial and the sample size J_i is sufficiently large in every group (for example, a lot of people live in Wyoming, even if not nearly as many as in California), the variance of the group-average error term may be well approximated by σ_c^2, which is homoskedastic. In that case, OLS applied to Equation 1 is nearly the best linear unbiased estimator. In contrast, if one weights by $\sqrt{J_i}$, as in Equation 2, the reweighted error term has variance $J_i\sigma_c^2 + \sigma_u^2$, which could be highly heteroskedastic. This provides an explanation for why weighting by the within-group sample size sometimes leads to less precise estimation than OLS.[9] On the other hand, if σ_c^2 is small and the within-group sample size J_i is highly variable and small in some groups, weighting by the within-group sample size may indeed improve the precision of estimation, sometimes by a great deal.

So what is a practitioner to do? Fortunately, as Dickens points out, it is easy to approach the heteroskedasticity issue as an empirical question. One way to go is to start with OLS estimation of Equation 1 and then use the OLS residuals to perform the standard heteroskedasticity diagnostics we teach in introductory econometrics. For example, in this situation, the modified Breusch-Pagan test described in Wooldridge (2013, pp. 276–78) comes down to just applying OLS to a simple regression of the squared OLS residuals on the inverse within-group sample size $1 / J_i$. The significance

[9] An important related point is that, if one has access to the individual-level data on y_{ij} and applies OLS to the regression of y_{ij} on X_i, this is numerically identical to the group-average WLS of Equation 2 and hence suffers from the same inefficiency associated with ignoring the clustered nature of the error term. For more discussion of the mapping between individual-level and group-average regressions, see Wooldridge (2003) and Donald and Lang (2007).

of the *t*-ratio for the coefficient on $1 / J_i$ indicates whether the OLS residuals display significant evidence of heteroskedasticity. The test therefore provides some guidance for whether weighted estimation seems necessary. A remarkable feature of this test is that the estimated intercept consistently estimates σ_c^2 and the estimated coefficient of $1 / J_i$ consistently estimates σ_u^2. This enables an approximation of the variance structure in Equation 4, which then can be used to devise a more refined weighting procedure that, unlike the simple weighting scheme in Equation 2, takes account of the group error component c_i.[10]

Of course, this particular heteroskedasticity test explores only the relationship of the error variance to the within-group sample size. It does not address heteroskedasticity related to the explanatory variables, but that also can be examined with standard diagnostics, such as those reviewed in Wooldridge (2013, Chapter 8). So our first recommendation to practitioners is not to assume that heteroskedasticity is (or is not) an issue but rather to perform appropriate diagnostics before deciding.

We wish to make two additional recommendations. One is that, regardless of whether and how one weights, the inevitable uncertainty about the precise nature of the true variance structure means that some heteroskedasticity may remain in the error term. We therefore recommend reporting heteroskedasticity-robust standard errors.

Also, it often is good practice to report both weighted and unweighted estimates. For one thing, as in our divorce example, a comparison of the robust standard errors is instructive about which estimator is more precise. But there is an additional consideration. Under exogenous sampling and correct specification of the conditional mean of *y* in Equation 1, both OLS and WLS are consistent for estimating the regression coefficients. On the other hand, under either the endogenous sampling discussed in the next section or model misspecification (an example of which is the failure to model heterogeneous effects, to be discussed in Section V), OLS and WLS generally have different probability limits. Therefore, as suggested by DuMouchel and Duncan (1983), the contrast between OLS and WLS estimates can be used as a diagnostic for model misspecification or endogenous sampling.[11]

In truth, of course, the parametric models we use for estimating causal effects are nearly always misspecified at least somewhat. Thus, the practical question is not whether a chosen specification is exactly the true data-generating process but rather whether it is a good enough approximation to enable nearly unbiased and consistent estimation of the causal effects of interest. When weighted and unweighted estimates contradict each other, this may be a red flag that the specification is not a good enough approximation to the true form of the conditional mean. For example, Lee and Solon (2011) find that, when the dependent variable used in the divorce-rate regressions is specified not in logs but in levels (as in both the Friedberg and Wolfers studies), the OLS and WLS estimates are dramatically different from each other. This in itself does not pinpoint exactly what is wrong with the linear-in-levels model specification, but it is a valuable warning sign that the issue of functional form specification warrants further attention.

[10] In particular, this feasible generalized least squares estimator would weight by the square root of the estimated inverse of the quantity in Equation 4 instead of by $\sqrt{J_i}$.

[11] See Deaton (1997, p. 72) for a clear exposition of how to assess the statistical significance of the contrast between OLS and WLS estimates of a linear regression model when, under the null hypothesis, OLS is efficient. For a more general treatment, see Wooldridge (2001, pp. 463–64).

IV. Correcting for Endogenous Sampling

An altogether different motive for weighting in research on causal effects is to achieve consistent estimation in the presence of endogenous sampling — that is, sampling in which the probability of selection varies with the dependent variable even after conditioning on the explanatory variables. A nice example comes from the classic paper on choice-based sampling by Manski and Lerman (1977). Suppose one is studying commuters' choice of transit mode, such as the choice between driving to work and taking the bus. One might be particularly interested in how certain explanatory variables, such as bus fare and walking distance to and from bus stops, affect the probability of choosing one mode versus the other. Given a random sample of commuters, most empirical researchers would perform maximum likelihood estimation of a probit or logit model for the binary choice between transit modes.

But suppose the sample is drawn not as a random sample of commuters but as a choice-based sample. As Manski and Lerman explain, "[I]n studying choice of mode for work trips, it is often less expensive to survey transit users at the station and auto users at the parking lot than to interview commuters at their homes." Manski and Lerman show that, if the resulting sample overrepresents one mode and underrepresents the other relative to the population distribution of choices, maximizing the conventional log likelihood (which is an incorrect log likelihood because it fails to account for the endogenous sampling) generally results in inconsistent parameter estimation.[12] And if instead one maximizes the quasi-log likelihood that weights each observation's contribution to the conventional log likelihood by its inverse probability of selection from the commuter population (thus using a reverse funhouse mirror to make the sample representative of the population), consistent estimation of the parameters is restored.

Another example is estimating the earnings return to an additional year of schooling. Most labor economists would frame their analysis within a linear regression of log earnings on years of schooling with controls for other variables such as years of work experience. Although that regression model has been estimated countless times by OLS, researchers cognizant of the endogeneity of years of schooling often have sought to devise instrumental variables (IV) estimators of the regression. In any case, if the regression were estimated with the full PSID without any correction for the oversampling of the low-income population, this would lead to inconsistent estimation of the regression parameters. The sampling would be endogenous because the sampling criterion, family income, is related to the error term in the regression for log earnings. Again, however, for an estimation strategy that would be consistent if applied to a representative sample, suitably weighted estimation would achieve consistency. For example, if the schooling variable somehow were exogenous so that OLS estimation with a representative sample would be consistent, then applying WLS to the endogenously selected sample (weighting each contribution to the sum of squares by its inverse probability of selection) also would be consistent. This could be achieved

[12] They also note a quirky exception: In a logit model that includes mode-specific intercepts in the associated random-utility model, the coefficients of the other explanatory variables are consistently estimated. That is a peculiar feature of the logit specification, and it does not carry over to other specifications such as the probit model. Furthermore, without a consistent estimate of the intercept, one cannot obtain consistent estimates of the average partial effects, which are commonly reported in empirical studies.

by applying least squares to an equation that looks like Equation 2, but now with J_i standing for the inverse probability of selection. Similarly, if one were to perform IV estimation, one would need to weight the IV orthogonality conditions by the inverse probabilities of selection.

These examples illustrate a more general point, analyzed formally in Wooldridge (1999) for the entire class of *M*-estimation. In the presence of endogenous sampling, estimation that ignores the endogenous sampling generally will be inconsistent. But if instead one weights the criterion function to be minimized (a sum of squares, a sum of absolute deviations, the negative of a log likelihood, a distance function for orthogonality conditions, etc.) by the inverse probabilities of selection, the estimation becomes consistent.

An important point stressed in Wooldridge (1999) is that, if the sampling probabilities vary exogenously instead of endogenously, weighting might be unnecessary for consistency and harmful for precision. In the case of a linear regression model that correctly specifies the conditional mean, the sampling would be exogenous if the sampling probabilities are independent of the error term in the regression equation. This would be the case, for example, if the sampling probabilities vary only on the basis of explanatory variables. More generally, the issue is whether the sampling is independent of the dependent variable conditional on the explanatory variables.

For example, suppose one estimates a linear regression model with a sample that overrepresents certain states (as in the Current Population Survey) but the model includes state dummy variables among the explanatory variables. Then, if the model is correctly specified (more about that soon), the error term is not related to the sampling criterion and weighting is unnecessary. If the error term obeys the ideal conditions, then OLS estimation is optimal in the usual way.[13] While weighting is unnecessary for consistency in this instance, is there any harm in using WLS instead of OLS? Yes, there can be an efficiency cost. If the error term was homoskedastic prior to weighting, the weighting will induce heteroskedasticity, with the usual consequence of imprecise estimation. More generally, when the error term may have been heteroskedastic to begin with, the efficiency comparison between weighting or not weighting by inverse probabilities of selection becomes less clear. Again, as in Section III, we recommend using standard diagnostics for heteroskedasticity as a guide in the search for an efficient estimator.

Of course, we need again to acknowledge that, in practice, one's model is almost never perfectly specified. At best, it is a good approximation to the data-generating process. As a result, just as theorems in microeconomic theory based on unrealistically strong assumptions provide only rough guidance about what is going on in the actual economy, theorems from theoretical econometrics provide inexact (though valuable,

[13] Another practical example is where the survey organization provides sampling weights to adjust for differential nonresponse, including attrition from a panel survey, and these weights are based only on observable characteristics that are controlled for in the regression model (perhaps gender, race, age, location). In this situation, it is not clear that there is an advantage to using such weights. Matters become particularly complicated when the true nonresponse probability also depends on unobservables. In this case, which is an instance of the sample-selection-bias situation famously analyzed by Heckman (1979), consistent coefficient estimation is often unattainable unless one is willing to rely on strong parametric assumptions. For more on correcting for nonresponse, see Wooldridge (2002) and Fitzgerald, Gottschalk, and Moffitt (1998).

in our view) guidance about how to do empirical research. In that light, reconsider the example in the previous paragraph. If the sampling probability varies only across states and the regression model that controls for state dummies is a good, though imperfect, approximation to the true model for the conditional mean, then one might reasonably hope that OLS estimation would come close to unbiased and consistent estimation of the effects of the explanatory variables. The same goes for WLS estimation (which also would fall short of perfect unbiasedness and consistency[14]), but WLS might be less precise.

In the end, what is our advice to practitioners? First, if the sampling rate varies endogenously, estimation weighted by the inverse probabilities of selection is needed on consistency grounds. Second, the weighted estimation should be accompanied by robust standard errors. For example, in the case of a linear regression model, the heteroskedasticity induced by the weighting calls for the use of White (1980) heteroskedasticity-robust standard errors.[15] Finally, when the variation in the sampling rate is exogenous, both weighted and unweighted estimation are consistent for the parameters of a correctly specified model, but unweighted estimation may be more precise. Even then, as in the previous section, we recommend reporting both the weighted and unweighted estimates because the contrast serves as a useful joint test against model misspecification and/or misunderstanding of the sampling process.

V. Identifying Average Partial Effects

To consider a third motivation for weighted estimation of causal effects, return to the example of divorce-rate regressions. Recall that, when Lee and Solon followed Friedberg and Wolfers in using the level, rather than the log, of the divorce rate as the dependent variable, they found that the OLS and WLS estimates differed dramatically from each other, with the WLS results showing more evidence of a positive impact of unilateral divorce on divorce rates. One possible explanation is that, if the impact of unilateral divorce is heterogeneous—that is, if it interacts with other state characteristics—then OLS and WLS estimates that do not explicitly account for those interactions may identify different averages of the heterogeneous effects. For example, if unilateral divorce tends to have larger effects in more populous states, then WLS estimation that places greater weight on more populous states

[14] In Footnote 17 in the next section, we will mention special cases of model misspecification where, although OLS may be inconsistent for estimating particular causal effects, certain weighted estimators do achieve consistency.

[15] Wooldridge (1999) presents the appropriate "sandwich" estimator of the asymptotic variance-covariance matrix for the general case of *M*-estimation under endogenous sampling. Wooldridge (2001) analyzes a subtly different sort of sampling called "standard stratified sampling." In standard stratified sampling, the survey selects a *deterministically* set number of observations per stratum. In this case, the asymptotic variance-covariance matrix is more complex and is strictly smaller than the one analyzed in Wooldridge (1999). Intuitively, sampling variability is reduced by not leaving the within-stratum sample sizes to chance. In the example of a linear regression model, the White heteroskedasticity-robust standard errors then become conservative in the sense that they are slightly upward-inconsistent estimates of the true standard deviations of the estimated coefficients.

will tend to estimate larger effects than OLS does. Indeed, Lee and Solon found that, when they redid WLS with California omitted from the sample, the estimated effects of unilateral divorce came out smaller and more similar to the OLS estimates, which gave the same weight to California as to any other state.

This raises the question of whether one might want to weight in order to identify a particular average of heterogeneous effects, such as the population average partial effect. Indeed, we have the impression that many empirical practitioners believe that, by performing WLS with weights designed to reflect population shares, they do achieve consistent estimation of population average partial effects (for example, the average impact of unilateral divorce on divorce rates for the U.S. population).[16] This belief may be based on the fact, discussed above in Section IIA, that this WLS approach does consistently estimate the population linear projection of the dependent variable on the explanatory variables. That, however, is not the same thing as identifying the population average partial effects.[17] For a previous demonstration of this point, see Deaton (1997, pp. 67–70).

Here, we illustrate with a simple cross-sectional example. Suppose the true model for an individual-level outcome y_i is

$$(5)\quad y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \beta_4 X_i D_i + v_i$$

where D is a dummy variable indicating urban (rather than rural) location and the error term v is fully independent of all the explanatory variables. Then the effect of X on y is heterogeneous, with β_2 as the rural effect and $\beta_2 + \beta_4$ as the urban effect. The average effect for the population is the population-weighted average of these two effects, which is $\beta_2 + \beta_4\pi$ where π represents the urban share of the population.

Suppose that one fails to model the heterogeneity of effects and instead estimates the regression of y on just X and D, with the interaction term omitted. And, to represent the situation in which the researcher is tempted to weight, suppose that one does so with data from a survey that oversampled in the urban sector, so that the urban fraction of the sample is $p > \pi$. The OLS estimator of the coefficient of X does identify a particular weighted average of the rural and urban effects, but no one would expect that weighted average to be the same as the population average effect. After all, the sample systematically overrepresents the urban sector. And, as we soon will show, that is indeed one of the reasons that the probability limit of the OLS estimator differs from the population average partial effect. But the math also will reveal a second reason. In least squares estimation, observations with extreme values of the explanatory variables have particularly large influence on the estimates. As a result, the weighted average of the rural and urban effects identified by OLS depends not only on the sample shares of the two sectors but also on how the within-sector variance of X differs between the two sectors.

[16] To cite one of many possible examples, Footnote 9 in Ludwig and Miller's (2007) deservedly influential regression-discontinuity evaluation of Head Start seems to make this claim in distinguishing the estimands when population weighting is or is not used.

[17] One exception in which it *is* the same thing is in a simple regression on one dummy regressor—that is, a simple contrast between the means for two subpopulations. And this exception extends to the case of a "fully saturated" regression on a set of category dummies, which is a contrast among means for multiple subpopulations. Another case in which using suitably weighted estimators to identify population linear projections identifies a population average causal effect is the "doubly robust" estimator of treatment effects introduced by Robins, Rotnitzky, and Zhao (1994) and analyzed by Wooldridge (2007).

Now suppose that instead one estimates the regression of y on X and D by WLS with weighting by the inverse probabilities of selection. By reweighting the sample to get the sectoral shares in line with the population shares, WLS eliminates the first reason that OLS fails to identify the population average partial effect, but it does not eliminate the second. As a result, the WLS estimator and the OLS estimator identify different weighted averages of the heterogeneous effects, and *neither one* identifies the population average effect.

To be precise, let $\hat{\beta}_{2,OLS}$ denote the OLS estimator of the coefficient of X when the interaction term is omitted, and let $\hat{\beta}_{2,WLS}$ denote the corresponding WLS estimator. It is straightforward to show that the probability limit of the latter is what one would get from the corresponding population linear projection:

$$(6)\quad \operatorname{plim}\hat{\beta}_{2,WLS} = \beta_2 + \beta_4\left[\frac{\pi\sigma_1^2}{\pi\sigma_1^2 + (1-\pi)\sigma_0^2}\right]$$

where σ_0^2 and σ_1^2 respectively denote the within-sector variances of X for the rural and urban sectors. In contrast, the probability limit of the OLS estimator is

$$(7)\quad \operatorname{plim}\hat{\beta}_{2,OLS} = \beta_2 + \beta_4\left[\frac{p\sigma_1^2}{p\sigma_1^2 + (1-p)\sigma_0^2}\right].$$

If the effect of X were homogeneous (that is, if $\beta_4 = 0$), then both estimators would be consistent for the homogeneous effect β_2. Which estimator is preferable would depend on which is more precise, the question we already discussed in Section III's analysis of heteroskedasticity.

The point of the present section, however, is to consider the heterogeneous-effects case where $\beta_4 \neq 0$. In that case, Equations 6 and 7 imply that the inconsistencies of the two estimators with respect to the true population average partial effect $\beta_2 + \beta_4\pi$ are

$$(8)\quad \operatorname{plim}\hat{\beta}_{2,WLS} - (\beta_2 + \beta_4\pi) = \beta_4\left[\frac{\pi\sigma_1^2}{\pi\sigma_1^2 + (1-\pi)\sigma_0^2} - \pi\right]$$

and

$$(9)\quad \operatorname{plim}\hat{\beta}_{2,OLS} - (\beta_2 + \beta_4\pi) = \beta_4\left[\frac{p\sigma_1^2}{p\sigma_1^2 + (1-p)\sigma_0^2} - \pi\right].$$

In the knife-edge special case where $\sigma_0^2 = \sigma_1^2$, WLS is consistent for the population average effect and OLS is not. More generally, though, *both* estimators are inconsistent for the population average effect (or any other average effect that researchers commonly consider interesting). With either over- or undersampling of the urban sector ($p \neq \pi$), WLS and OLS are inconsistent in different ways and neither strictly dominates the other. It is easy to concoct examples in which each is subject to smaller inconsistency than the other.

Here are the lessons we draw from this example. First, we urge practitioners not to fall prey to the fallacy that, in the presence of unmodeled heterogeneous effects, weighting to reflect population shares generally identifies the population average partial effect.[18] Second, we reiterate the usefulness of the contrast between weighted and unweighted estimates. We said before that the contrast can serve as a test for misspecification, and the failure to model heterogeneous effects is one sort of misspecification that can generate a significant contrast. Third, where heterogeneous effects are salient, we urge researchers to *study* the heterogeneity, not just try to average it out. Typically, the average partial effect is not the only quantity of interest, and understanding the heterogeneity of effects is important. For example, unless one understands the heterogeneity, it is impossible to extrapolate from even a well-estimated population average effect in one setting to what the average effect might be in a different setting. In the simple example above, this recommendation just amounts to advising the practitioner to include the interaction term instead of omitting it. We understand that, in most empirical studies, studying the heterogeneity is more complex, but we still consider it worthwhile.

VI. Summary and General Recommendations for Empirical Practice

In Section II, we distinguished between two types of empirical research: (1) research directed at estimating population descriptive statistics, and (2) research directed at estimating causal effects. For the former, weighting is called for when it is needed to make the analysis sample representative of the target population. For the latter, the question of whether and how to weight is more nuanced.

In Sections III, IV, and V, we proceeded to discuss three distinct potential motives for weighting when estimating causal effects: (1) to achieve more precise estimates by correcting for heteroskedasticity; (2) to achieve consistent estimates by correcting for endogenous sampling; and (3) to identify average partial effects in the presence of unmodeled heterogeneity of effects. In our detailed discussion of each case, we have noted instances in which weighting is not as good an idea as empirical researchers sometimes think. Our overarching recommendation therefore is to take seriously the question in our title: What are we weighting for? Be clear about the reason that you are considering weighted estimation, think carefully about whether the reason really applies, and double-check with appropriate diagnostics.

A couple of other recurring themes also bear repeating. In situations in which you might be inclined to weight, it often is useful to report both weighted and unweighted estimates and to discuss what the contrast implies for the interpretation of the results. And, in many of the situations we have discussed, it is advisable to use robust standard errors.

[18] This raises the question of when and how one *can* identify the population average effect. The answer depends on the particular circumstances. Returning to the divorce example, suppose that the effect of unilateral divorce on the divorce rate varied across states. If one could estimate a state-specific effect for each of the 50 states, then one could estimate the population average effect by taking a population-weighted average of the 50 state-specific estimates. In fact, however, one could not identify state-specific effects for those states that did not switch to unilateral divorce during the sample period unless one added further assumptions about how the effect heterogeneity is related to observable variables.

References

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

Autor, David H., Lawrence F. Katz, and Alan B. Krueger. 1998. "Computing Inequality: Have Computers Changed the Labor Market?" *Quarterly Journal of Economics* 113(4):1169–213.

Borjas, George J. 2003. "The Labor Demand Curve Is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market." *Quarterly Journal of Economics* 118(4):1335–74.

Card, David, and Alan B. Krueger. 1992. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy* 100(1):1–40.

Deaton, Angus. 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore: The Johns Hopkins University Press.

Dehejia, Rajeev, and Adriana Lleras-Muney. 2004. "Booms, Busts, and Babies' Health." *Quarterly Journal of Economics* 119(3):1091–130.

Dickens, William T. 1990. "Error Components in Grouped Data: Is It Ever Worth Weighting?" *Review of Economics and Statistics* 72(2):328–33.

Donald, Stephen G., and Kevin Lang. 2007. "Inference with Difference-in-Differences and Other Panel Data." *Review of Economics and Statistics* 89(2):221–33.

Donohue, John J., III, and Steven D. Levitt. 2001. "The Impact of Legalized Abortion on Crime." *Quarterly Journal of Economics* 116(2):379–420.

DuMouchel, William H., and Greg J. Duncan. 1983. "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples." *Journal of the American Statistical Association* 78(383):535–43.

Elder, Todd E., John H. Goddeeris, and Steven J. Haider. 2011. "A Deadly Disparity: A Unified Assessment of the Black-White Infant Mortality Gap." *B.E. Journal of Economic Analysis and Policy (Contributions)* 11(1):Article 33.

Fitzgerald, John, Peter Gottschalk, and Robert Moffitt. 1998. "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics." *Journal of Human Resources* 33(2):251–99.

Friedberg, Leora. 1998. "Did Unilateral Divorce Raise Divorce Rates? Evidence from Panel Data." *American Economic Review* 88(3):608–27.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1):153–61.

Lee, Jin Young, and Gary Solon. 2011. "The Fragility of Estimated Effects of Unilateral Divorce Laws on Divorce Rates." *B.E. Journal of Economic Analysis and Policy (Contributions)* 11(1):Article 49.

Levitt, Steven D. 1998. "Juvenile Crime and Punishment." *Journal of Political Economy* 102(4):1156–85.

Ludwig, Jens, and Douglas L. Miller. 2007. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." *Quarterly Journal of Economics* 122(1):159–208.

Manski, Charles F., and Steven R. Lerman. 1977. "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica* 45(8):1977–88.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89(427):846–66.

Shin, Donggyun, and Gary Solon. 2011. "Trends in Men's Earnings Volatility: What Does the Panel Study of Income Dynamics Show?" *Journal of Public Economics* 95(7–8):973–82.

U.S. Bureau of the Census. 1968. *Current Population Reports: Consumer Income*, P-60(55). Washington, D.C.: U.S. Government Printing Office.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48(4):817–38.

Wolfers, Justin. 2006. "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results." *American Economic Review* 96(5):1802–20.

Wooldridge, Jeffrey M. 1999. "Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples." *Econometrica* 67(6):1385–406.

———. 2001. "Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples." *Econometric Theory* 17(2):451–70.

———. 2002. "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition, and Stratification." *Portuguese Economic Journal* 1(2):117–39.

———. 2003. "Cluster-Sample Methods in Applied Econometrics." *American Economic Review* 93(2):133–38.

———. 2007. "Inverse Probability Weighted Estimation for General Missing Data Problems." *Journal of Econometrics* 141(2):1281–301.

———. 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd edition. Cambridge: MIT Press.

———. 2013. *Introductory Econometrics: A Modern Approach*, 5th edition. Mason: South-Western.