*Regressions can be weighted by propensity scores in order to reduce bias. However, weighting is likely to increase random error in the estimates, and to bias the estimated standard errors downward, even when selection mechanisms are well understood. Moreover, in some cases, weighting will increase the bias in estimated causal parameters. If investigators have a good causal model, it seems better just to fit the model without weights. If the causal model is improperly specified, there can be significant problems in retrieving the situation by weighting, although weighting may help under some circumstances.*

*Keywords: causation, selection, models, experiments, observational studies, regression, propensity scores*

# Weighting Regressions by Propensity Scores

DAVID A. FREEDMAN
*University of California, Berkeley*

RICHARD A. BERK
*University of Pennsylvania*

Estimating causal effects is often the key to evaluating social programs, but the interventions of interest are seldom assigned at random. Observational data are therefore frequently encountered. In order to estimate causal effects from observational data, some researchers weight regressions using "propensity scores." This simple and ingenious idea is due to Robins and his collaborators. If the conditions are right, propensity scores can be used to advantage when estimating causal effects.

However, weighting has been applied in many different contexts. The costs of misapplying the technique, in terms of bias and variance, can be serious. Many users, particularly in the social sciences, seem unaware of the pitfalls. Therefore, it may be useful to explain the idea and the circumstances under which it can go astray.

That is what we try to do here. We illustrate the performance of the technique—and some of the problems that can arise—on simulated data where the causal mechanism and the selection mechanism are both known, which makes it easy to calibrate performance.

AUTHORS' NOTE: We would like to thank Larry Brown, Rob Gould, Rob Hollister, Guido Imbens, Brian Kriegler, Dan Nagin, Jamie Robins, Paul Rosenbaum, Dylan Small, Mikhail Traskin, David Weisburd, and Peter Westfall for many helpful comments. Any remaining infelicities are the responsibility of the authors. Richard Berk's work on this paper was funded in part by a grant from the National Science Foundation: SES-0437169, "Ensemble methods for Data Analysis in the Behavioral, Social and Economic Sciences." DRAFT 8/10/2008 System time 951

We focus on cross-sectional parametric models, of the kind commonly seen in applications. Pooling time-series and cross-sectional variation leads to substantial additional complexity. Thus, we consider linear causal models like

$$Y = a + bX + c_1 Z_1 + c_2 Z_2 + U, \tag{1}$$

where $X = 1$ or 0 according as the subject is in treatment or control; $Z_1$ and $Z_2$ are confounders, correlated with $X$. The random error $U$ is independent of $X$, $Z_1$, and $Z_2$.

The "propensity score" $\hat{p}$ is an estimate for $P(X = 1|Z_1, Z_2)$, that is, the conditional probability of finding the subject in the treatment group given the confounders. Subjects with $X = 1$ receive weight $1/\hat{p}$; subjects with $X = 0$ receive weight $1/(1 - \hat{p})$. A "weighted" regression minimizes the weighted sum of squares.

We investigated the operating characteristics of weighting in a dozen simulation models. In these simulations, there were $n = 1000$ independent, identically distributed subjects. In some cases, we reran the simulation with $n = 10{,}000$ subjects, to see the effect of larger $n$ on bias and variance.

Each simulation had two components. The first component was a model that explained selection into the treatment or control condition. The second component was a causal model that determined response to treatment and to confounders. (Responses may be continuous or binary.) Selection was exogenous, that is, independent of the error term in the causal model.

The simulations were all favorable to weighting, in three important ways: (i) subjects were independent and identically distributed, (ii) selection was exogenous, and (iii) the selection equation was properly specified. We report in detail on two simulations that were reasonably typical, and mention some others in passing. We write $Y$ for the response, $X$ for treatment status (0 if in control, 1 if in treatment), and $Z$ for the confounder. Generally, $Z$ is multivariate normal.

## 1. SIMULATION #1

Our first simulation had a continuous linear response and probit selection. The causal model is

$$Y = a + bX + c_1 Z_1 + c_2 Z_2 + dU, \tag{2}$$

where $U$ is $N(0, 1)$. The selection model is

$$X = (e + f_1 Z_1 + f_2 Z_2 + V > 0), \tag{3}$$

where $V$ is $N(0, 1)$. Here, $a, b, c_1, c_2, d, e, f_1$, and $f_2$ are parameters.

Equation (3) may look a bit cryptic. More explicitly, the equation says that $X = 1$ if $e + f_1 Z_1 + f_2 Z_2 + V > 0$; otherwise, $X = 0$.

By construction, $U$, $V$, and $Z = (Z_1, Z_2)$ are all independent, and $Z$ is bivariate normal. The observables are $(X, Z, Y)$. The variables $U$ and $V$ are not observable. In particular, $X$ follows a probit model. To construct the weights, we fit this probit model to the data on $(X, Z)$.

Let $\hat{p}$ be the estimated probability that $X = 1$ given $Z$. Subjects with $X = 1$ get weight $w = 1/\hat{p}$. Subjects with $X = 0$ get weight $w = 1/(1-\hat{p})$. Notice that $\hat{p}$ depends on $Z$, so $w$ depends on $X$ and $Z$. Notice too that the selection equation is correctly specified.

For simplicity, we put $a = b = c_1 = d = 1$ and $c_2 = 2$ in equation (2). To keep variability in the weights within bounds, we make $e = .5$, $f_1 = .25$, and $f_2 = .75$ in equation (3). We set $\text{var}(Z_1) = 2$, $\text{var}(Z_2) = 1$, $\text{cov}(Z_1, Z_2) = 1$, $E(Z_1) = .5$, and $E(Z_2) = 1$.

We run regressions of $Y$ on $X$ and $Z$, unweighted and weighted, getting estimates for $a, b, \ldots,$ and their nominal standard errors. ("Nominal" standard errors are computed from the usual regression formulae.) We also run a regression of $Y$ on $X$ and $Z_1$. Finally, we run a simple regression of $Y$ on $X$.

Without the weights, the latter two regressions are misspecified: there is omitted-variables bias. The point of the weighting, as in most of the social-science literature we reviewed, is to correct omitted-variables bias. In the simulations, truth is known, so we can evaluate the extent to which the correction succeeds.

We repeat the process 250 times, getting the mean of the estimates, the standard deviation of the estimates, and the root mean square of the nominal standard errors. We abbreviate SD for standard deviation, SE for standard error, and RMS for root mean square. The SD measures the likely size of the random error in the estimates.

If $Z_1$ and $Z_2$ are both included in the regression, the weighted multiple regression estimates are essentially unbiased. However, the SD of the $\hat{b}$'s is about double the SD in the unweighted regression. Furthermore, the nominal SEs are too small by a factor of three (Table 1, first two blocks). When all the covariates are included, weighting the regression is therefore counterproductive. There is no bias to reduce, there is an increase in variance, and the nominal SEs become difficult to interpret.

Next, suppose $Z_2$ is omitted from the regression. The unweighted regression of $Y$ on $X$ and $Z_1$ then gives a biased estimate for $b$. The weighted regression of $Y$ on $X$ and $Z_1$ is still somewhat biased for $b$, and quite biased for $a$ and $c_1$. The bias in $\hat{b}$ is "small-sample bias." The other biases will not disappear with larger samples. The SDs in the weighted

TABLE 1. Simulation #1. Linear regression with $n = 1000$ independent subjects. "Ave" is the average value of the estimates and "SD" is their standard deviation, across 250 replications. "nom SE" is the nominal SE. The table reports the RMS of the nominal SEs.

| | Parameters | | | |
|---|---|---|---|---|
| | $a$ | $b$ | $c_1$ | $c_2$ |
| True values | 1 | 1 | 1 | 2 |
| Linear regression of $Y$ on $X$, $Z_1$ and $Z_2$, unweighted | | | | |
| Ave | 0.9970 | 1.0101 | 1.0003 | 1.9952 |
| SD | 0.0802 | 0.0974 | 0.0323 | 0.0468 |
| nom SE | 0.0812 | 0.0967 | 0.0320 | 0.0466 |
| Linear regression of $Y$ on $X$, $Z_1$ and $Z_2$, weighted | | | | |
| Ave | 1.0007 | 1.0089 | 0.9947 | 1.9978 |
| SD | 0.1452 | 0.2130 | 0.1010 | 0.1400 |
| nom SE | 0.0562 | 0.0635 | 0.0320 | 0.0459 |
| Linear regression of $Y$ on $X$ and $Z_1$, unweighted | | | | |
| Ave | 1.6207 | 2.1310 | 1.8788 | |
| SD | 0.1325 | 0.1574 | 0.0446 | |
| nom SE | 0.1345 | 0.1569 | 0.0415 | |
| Linear regression of $Y$ on $X$ and $Z_1$, weighted | | | | |
| Ave | 2.3994 | 1.1366 | 1.9432 | |
| SD | 0.2995 | 0.3295 | 0.1202 | |
| nom SE | 0.0789 | 0.1082 | 0.0401 | |
| Linear regression of $Y$ on $X$, unweighted | | | | |
| Ave | 0.1547 | 5.0232 | | |
| SD | 1.1101 | 1.0830 | | |
| nom SE | 0.2276 | 0.2495 | | |
| Linear regression of $Y$ on $X$, weighted | | | | |
| Ave | 3.0665 | 1.4507 | | |
| SD | 0.7880 | 0.7765 | | |
| nom SE | 0.1414 | 0.1972 | | |

regression are rather large, and the nominal SEs are too small (Table 1, middle two blocks).

Finally, suppose $Z_1$ and $Z_2$ are both omitted from the regression. The bias in the weighted regression is even worse. By comparison, an unweighted simple regression does better at estimating $a$, worse at estimating $b$ (Table 1, last two blocks). Again, the bias in the weighted regression estimate for $b$ is a small-sample bias: with an $n$ of 10,000, this bias will largely disappear.

The bias in $\hat{a}$ comes about because $E(Z) \neq 0$. This bias remains no matter how large the sample may be. If we wish to estimate the causal effects of the treatment and control regimes separately, conditional on the covariates, this bias cannot be ignored. (It does cancel if we estimate differential effects.)

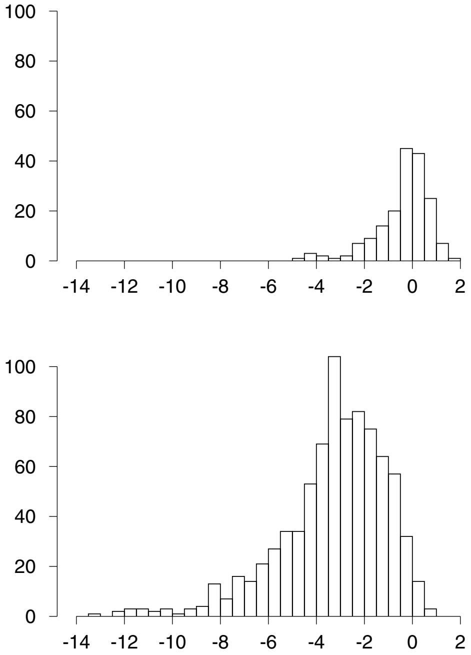
Some of the trouble is due to variability in the weights. We did the simulation over again, truncating the weights at 20: in other words, when the weight is above 20, we replace it by 20. Qualitatively, results are similar. Quantitatively, there is a noticeable reduction in variance—even though we only trim 6 weights per 1000 subjects. However, there is some increase in bias. We also tried filtering out subjects with large weights. This was worse than truncation. Variability in the weights is a difficulty that is frequently encountered in applications.

The unweighted simple regression of $Y$ on $X$ has substantial bias, and the nominal SEs are far too optimistic. Why? The error term in this regression is $c_1 Z_1 + c_2 Z_2$. Some of this will be picked up in the intercept and the coefficient of $X$, explaining the bias. The remainder is heteroscedastic, partly because $X$ is a binary variable so $(X, Z_1, Z_2)$ cannot be jointly normal, partly because weighting converts homoscedastic errors to heteroscedastic. That explains the deficiencies in the nominal SEs.

We return to the weighted regressions. It seems natural to try the Huber-White correction, but this is unlikely to help. With omitted variables, errors do not have conditional expectation 0 given the included variables, even after we subtract the projection of the error vector onto the regressors. Again, $(X, Z_1, Z_2)$ isn't normal, and the projection operator depends on the weights. The key assumption behind the correction is false. (Outliers are another problem.) Indeed, the Huber-White correction did not work very well for us, even in the full multivariate regression. The reason for this last failure may be the length of the tail in the distribution of $1/\hat{p}$, which is our next topic.

Recall that the weights $w$ are defined as follows: $w = 1/\hat{p}$ for subjects with $X = 1$ and $w = 1/(1 - \hat{p})$ for subjects with $X = 0$, where $\hat{p}$ is the estimated value for $P(X = 1|Z_1, Z_2)$. A histogram for $\log \log w$ in

Figure 1. Top panel: weights for controls. Bottom panel: weights for treatment group. Log log transformation.



one replication is shown in Figure 1. The top panel shows the histogram for $X = 0$; the bottom panel, for $X = 1$.

The height of each bar shows the number of observations falling in the corresponding class interval; there were 180 observations with $X = 0$ and 820 with $X = 1$. That is why the bottom histogram is bigger. It also has longer tails. The difference in the length of the tails in the two distributions is one of the problems faced by the weighting procedure. (The difference is not due to the difference in sample sizes.)

The two logs are needed to get a decent-looking histogram. The low end of the scale corresponds to weights just above 1, that is, $\hat{p}$'s just below 1. The high end of the scale corresponds to weights on the order of 50 to 250 for $X = 0$, and 5 to 15 for $X = 1$, depending on how the random numbers fall. For the particular replication reported here, the maximal weights were about 150 and 7, respectively. However, maxima are notoriously vulnerable to chance fluctuations, and larger weights do occur.

Which way do our assumptions cut? The assumption that subjects are independent and identically distributed is favorable to the modeling enterprise. So is the exogeneity of the selection mechanism. Making $V$ normal is another kindness; without it, the selection equation would be misspecified. Making $U$ normal also seems to be generous, since the response equation is estimated by least squares.

Assuming $Z$ to be normal presents tradeoffs that are more complicated. With shorter-tailed distributions, weighting may work better. With longer-tailed distributions, which seem more common in practice, weighting is likely to do worse.

In our simulations, the exogenous regressors $Z_1$ and $Z_2$ are randomized afresh on each of the 250 repetitions. Generating the $Z$'s once and for all at the beginning and re-using the same $Z$'s throughout makes almost no difference to the results. (We tried it.) In principle, the SDs should go down a little, but the difference is too small to see.

## 2. RESULTS FOR SIMULATION #2

Simulation #2 is just like simulation #1, with logit selection and logit response; the parameter values remain the same, along with the joint distribution of $(Z_1, Z_2)$. The causal model is

$$Y = (a + bX + c_1Z_1 + c_2Z_2 + U > 0), \tag{4}$$

and the selection model is

$$X = (e + f_1Z_1 + f_2Z_2 + V > 0), \tag{5}$$

where $(Z_1, Z_2)$, $U$, and $V$ are independent; $U$ and $V$ follow the standard logistic distribution.

Results are much like those in Simulation #1. See Table 2. However, with omitted variables, the weighted logistic regression performs very poorly at estimating the coefficient $b$ of the treatment variable. (A "weighted" logistic regression maximizes the weighted log likelihood function.) When $Z_1$ and $Z_2$ are both omitted, the sign of $\hat{b}$ is usually wrong. The unweighted simple logistic regression does substantially better.

The bad behavior of the weighted simple logistic regression is not a small-sample problem. It is quite reproducible. We think it is due to occasional large weights. However, if we truncate the weights above at 20, there is no improvement in the weighted estimator. At 10—and this affects only 65/1000 of the weights—$\hat{b}$ has a fair chance of being positive. In practice, of course, it might be hard to tell how much truncation to do. We return to this point later.

TABLE 2. Simulation #2. Logistic regression with $n = 1000$ independent subjects. "Ave" is the average value of the estimates, and "SD" is their standard deviation, across 250 replications. "nom SE" is the nominal SE. The table reports the RMS of the nominal SEs.

| | Parameters | | | |
|---|---|---|---|---|
| | $a$ | $b$ | $c_1$ | $c_2$ |
| True values | 1 | 1 | 1 | 2 |
| Logistic regression of $Y$ on $X$, $Z_1$ and $Z_2$, unweighted | | | | |
| Ave | 1.0100 | 1.0262 | 1.0210 | 2.0170 |
| SD | 0.2372 | 0.2919 | 0.1611 | 0.2674 |
| nom SE | 0.2296 | 0.2750 | 0.1589 | 0.2525 |
| Logistic regression of $Y$ on $X$, $Z_1$ and $Z_2$, weighted | | | | |
| Ave | 1.0178 | 1.0616 | 1.0470 | 2.1018 |
| SD | 0.3084 | 0.3066 | 0.2593 | 0.4197 |
| nom SE | 0.1286 | 0.1943 | 0.0960 | 0.1453 |
| Logistic regression of $Y$ on $X$ and $Z_1$, unweighted | | | | |
| Ave | 1.5879 | 1.3711 | 1.5491 | |
| SD | 0.2140 | 0.2543 | 0.1396 | |
| nom SE | 0.2027 | 0.2452 | 0.1389 | |
| Logistic regression of $Y$ on $X$ and $Z_1$, weighted | | | | |
| Ave | 2.5934 | 0.3214 | 1.8977 | |
| SD | 0.3419 | 0.3218 | 0.2391 | |
| nom SE | 0.0977 | 0.1684 | 0.0788 | |
| Logistic regression of $Y$ on $X$, unweighted | | | | |
| Ave | 0.6779 | 1.9893 | | |
| SD | 1.1458 | 1.1778 | | |
| nom SE | 0.1367 | 0.2016 | | |
| Logistic regression of $Y$ on $X$, weighted | | | | |
| Ave | 3.9154 | −2.1168 | | |
| SD | 0.9632 | 0.9725 | | |
| nom SE | 0.0729 | 0.1190 | | |

## 3. COVARIATE BALANCE

Covariate balance in a sample after weighting is sometimes used to justify the results of propensity score weighted regression. We tried Simulation #1 with one covariate instead of two and slightly different values for the parameters $a$, $b$, . . . . About 40% of the time, the covariate balanced across treatment and control groups. In these data sets, the simple weighted regression estimator was nearly unbiased for $b$. But the SD of the $\hat{b}$'s was about double the SD in the unweighted multiple regression, and the nominal SE was much too small. Therefore, covariate balance in the data does not answer our arguments. In our setup, you are better off just running the unweighted multiple regression. Of course, the response equation is correctly specified, which counsels against weighting. The selection equation is correct too, but this counsels in favor of weighting.

## 4. DISCUSSION

When a linear causal model is correctly specified, weighting is counterproductive, because there is no bias to remove. On the other hand, when the model omits relevant variables, weighting regressions by propensity scores is worth considering. If the propensity scores can be accurately estimated, weighting may lead to a substantial reduction in bias—although, with realistic samples sizes, the bias that remains can be appreciable. The price of bias reduction is an increase in random error, along with a downward bias in the nominal SEs. See Table 1.

There are two threshold questions. (i) Were relevant variables omitted from the causal model? (ii) Is there enough information to estimate the propensity scores with good accuracy? If the answer to both questions is "yes," the propensity scores are likely to help reduce bias. However, the conjunction is improbable. If variables are missing from the causal model, variables are likely to be missing from the selection model too. In all our simulation models, the selection model was correctly specified, shifting the balance in favor of weighting.

When the response model is logit, weighting creates substantial bias in coefficient estimates. See Table 2. There are parameters that can usefully be estimated in a weighted logit specification, but these are not the usual parameters of interest. Similar comments apply to the probit model and the proportional hazards model. On the latter, see Hernán, Brumback, and Robins (2001).

In the simulations reported here, as in many social-science papers, weighting is not intended to correct specification errors other than omitted-variables bias. The errors we have in mind include heteroscedasticity, dependence between subjects, endogeneity (selection into treatment correlated with the error term in the causal model), and so forth. In some of our simulations, weighting worsens endogeneity bias in multiple regression, but helps in simple regression.

With non-parametric models for response and selection—this is closer to Robins' original conception—the issues will be different. Still, you need to get at least one of the two models (and preferably both) nearly right in order for weighting to help much. If both models are wrong, weighting could easily be a dead end. There are papers suggesting that under some circumstances, estimating a shaky causal model and a shaky selection model should be doubly robust. Our results indicate that under other circumstances, the technique is doubly frail.

Robins and his collaborators were not estimating structural equations. They were estimating contrasts: what would happen if you put everyone into the treatment condition? the control condition? This is not a suggestion to replace structural equations by non-parametric modeling and contrasts. Our point is that caution is needed when using new techniques. Sometimes you do have to read the fine print. Non-parametric models, Robins' work, and contrasts versus structural equations will be discussed below.

The bottom line for social scientists is this. If you have a causal model that you believe, you should probably just fit it to the data. If there are omitted variables but the propensity scores can somehow be estimated with reasonable accuracy, weighting the regression should reduce bias. If you believe the propensity scores but not the causal model, a good option might be weighted contrasts between the treatment and control groups. On the other hand, weighting is likely to increase random error by a substantial amount, and nominal standard errors (the ones printed out by the software) can be much too small.

If you are going to weight, it rarely makes sense to use the same set of covariates in the response equation and the selection equation. Furthermore, you should always look at the weights. If results are sensitive to a few large weights, it is time to reconsider. Finally, if you go beyond continuous response variables and weighted least squares, each combination of response model and fitting procedure has to be considered separately—to see what the weighted regression is going to estimate.

## 5. LITERATURE REVIEW

There have recently been a number of studies that apply propensity score weighting to causal models. Much of the research addresses topics of interest to social scientists. The studies proceed in two steps, which are mimicked by our simulations.

*Step 1.* A model (typically logit or probit) is used to estimate the probability of selection into the treatment and control groups. The treatment may be an explicit intervention such as hospice care (Gozalo and Miller, 2007). Or, it may reflect some feature of an ongoing social process, such as marriage (Sampson et al., 2006). The units of analysis may be individuals (Francesconi and Nicoletti, 2006), or larger entities such as neighborhoods (Tita and Ridgeway, 2007).

*Step 2.* Estimated probabilities from the first step are used to construct weights. The weights are then used to fit the causal model of substantive interest. The causal model can take a variety of forms: conventional linear regression (Francisco and Nicoletti, 2006), logistic regression (Bluthenthal et al., 2006), Poisson regression (Tita and Ridgeway, 2007), hierarchical Poisson regression (Sampson et al., 2006), or proportional hazards (McNiel and Binder, 2007).

Sample sizes generally range from several hundred to several thousand. There will typically be several dozen covariates. In one example (Schonlau, 2006), there were over 100 possible covariates to choose from, and the sample size was around 650.

Investigators differ on procedures used for choosing regressors in the causal model. Sometimes, all available covariates are used (McNiel and Binder, 2007). Sometimes there is a screening process, so that only variables identified as important or out of balance are included (Ridgeway, McCaffrey, and Morral, 2006). Typically, a multivariate model is used; sometimes, however, there are no covariates (Leslie and Theibaud, 2007).

Some investigators use rather elaborate estimation procedures, including the lasso (Ridgeway, McCaffrey, and Morral, 2006) and boosting (Schonlau, 2006). These estimation procedures, like the variable selection procedures and choice of response model—when combined with weighting—can change the meaning of the parameters that are being estimated. Thus, caution is in order.

Investigators may combine "robust" standard errors and non-linear response models like hierarchical Poisson regressions (Sampson et al., 2006). The use of robust standard errors implicitly acknowledges that the model has the wrong functional form (Freedman, 2006). However, specification error is rarely considered to be a problem.

In this literature, important details of the model specification often remain opaque. See, for instance, pp. 483–9 in Sampson et al. (2006): although the selection model is clear, the response model remains unclear.

Few authors consider the bias in nominal standard errors, or the problems created by large weights. We saw no mention of definitional problems created by nonlinear response models or complex estimation procedures.

Lunceford and Davidian (2004) summarize the theory of weighted regressions, with some informative simulations. However, the limitations of the technique are not fully described.

In a biomedical application, Hirano and Imbens (2001) recommend including interactions between the treatment dummy and the covariates. In our simulations, this sometimes reduced bias in the estimated intercept, but usually had little effect.

Two journals have special issues that explore the merits of propensity scores. This includes use of propensity scores in weighted regression and in earlier techniques, such as (i) creating match sets or (ii) computing weighted contrasts between treatment and control groups. See

*Review of Economics and Statistics*, February 2004, vol. 86, no. 1;
*Journal of Econometrics*, March-April 2005, vol.125, no. 1–2.

Other references of interest include Arceneaux, Gerber, and Green (2006), Glazerman, Levy, and Myers (2003), Peikes, Moreno, and Orzol (2008), Wilde and Hollister (2007). These authors point to serious weaknesses in the propensity-score methods that have been used for program evaluation.

The basic papers on weighted regression include Robins and Rotnitzky (1992, 1995), Robins, Rotnitzky, and Zhao (1994), Rotnitzky, Robins, and Scharfstein (1998), Bang and Robins (2005). The last describes simulations that show the power of weighted regressions when the assumptions behind the technique are satisfied, even approximately. Kang and Schafer (2007) criticize use of weighted regressions, a central issue being variability in the weights. There is a reply by Robins, Sued, Lei-Gomez, and Rotnitzky (2007). Also see Crump, Hotz, Imbens, and Mitnik (2007) on handling variable weights. Freedman (2008) describes a measure-theoretic justification for weighting, in terms of Radon-Nikodym derivatives.

Weighted regression should be distinguished from the methods suggested by Heckman (1978, 1979). For instance, if $U$ and $V$ in (2)–(3) are correlated, Heckman recommended maximum likelihood, or—in the linear case—including an additional term in the regression to center the errors.

When unbiased estimators do not exist, there are theorems showing that reduction in bias is generally offset by an increase in variance (Doss and Sethuraman, 1989). Evans and Stark (2002) provide a broader context for this discussion.

## 6. THEORY

Suppose we have a linear causal model as in Simulation #1,

$$Y = a + bX + c_1 Z_1 + c_2 Z_2 + dU, \tag{6}$$

where $(Z_1, Z_2)$ is correlated with $X$. However, we omit $Z_1$ and $Z_2$ when we run the regression. Omitted-variables bias is the consequence, and the regression estimator is inconsistent. If we weight the regression using propensity weights, then $Z_1$ and $Z_2$ will be asymptotically balanced between treatment ($X = 1$) and control ($X = 0$). In other words, after weighting, covariates will be independent of treatment status, and hence cannot confound the causal relationship.

From this perspective, what can we say about $\hat{a}$ in a weighted simple regression? (See Table 1, last block.) It turns out that $\hat{a}$ estimates, not $a$ itself, but $a + E(c_1 Z_1 + c_2 Z_2)$, which is the average effect of the control condition—averaged across all values of the confounders. Weighting changed the meaning of the estimand. This is often the case.

The discussion here is intended only as a useful heuristic, rather than rigorous mathematics. A rigorous treatment would impose moment conditions on weighted variables, distinguishing between estimated weights and true weights.

Theoretical treatments of weighted regression generally assume that subjects are independent and identically distributed (IID). This is a very strong assumption. By comparison, with structural models, the exogenous variables need not be independent or identically distributed across subjects. Instead, it is commonplace to condition on such variables.

The stochastic elements that remain are the latent variables in the selection and response equations. To be sure, if the latents in the two equations fail to be independent within subject, or fail to be IID across subjects, the models will be misspecified. With nonparametric models, the IID assumption may go deeper. That is our next topic.

## 7. NONPARAMETRIC ESTIMATION

Suppose subject $i$ is observed for time $t = 0, 1, 2, \ldots$. Subjects are assumed to be IID. In period $t > 0$, subject $i$ chooses to be in treatment ($X_{it} = 1$) or control ($X_{it} = 0$). This choice depends on a vector of covariates $Z_{it-1}$ defined in the previous period. There is a response $Y_{it}$ that depends on the choice of regime $X_{it}$ and on the covariates $Z_{it-1}$. Furthermore, $Z_{it}$ depends on $Z_{it-1}$, $X_{it}$, and $Y_{it}$. The functions $f$, $g$, and $h$ determine choice, response, and evolution of covariates respectively. These functions are unknown in form, although subject to a priori smoothness conditions. We do not allow them to depend on $i$ or $t$. There are unobserved random errors $U_{it}$, $V_{it}$, and $W_{it}$. These are assumed to be independent within subject and IID across subjects, with

$$X_{it} = f\big(Z_{it-1}, U_{it}\big), \tag{7a}$$

$$Y_{it} = g\big(Z_{it-1}, X_{it}\big) + V_{it}, \tag{7b}$$

$$Z_{it} = h\big(Z_{it-1}, X_{it}, Y_{it}\big) + W_{it}. \tag{7c}$$

The system is assumed to be complete: apart from the random errors, there are no unobserved covariates that influence treatment choice or response. (Social-science applications discussed above do not satisfy the completeness assumption—far from it.)

This is a rather complex environment, in which parametric models might not do very well. It is for this sort of environment that Robins and his colleagues developed weighting. The object was to determine what would happen if the choice equation (7a) was no longer operative, and various treatment regimes were imposed on the subjects—without changing the response functions $g$, $h$, or the random errors—a prospect that makes little sense in social-science applications like Sampson et al. (2006) or Schonlau (2006). Sampson et al. at least have the sort of longitudinal data structure where parametric models might run into trouble. Schonlau, among others, uses weights in a cross-sectional data structure.

## 8. CONTRASTS

Let $i$ index the subjects in $T$ and $j$ index the subjects in $C$, so $w_i = 1/\hat{p}_i$ and $w_j = 1/(1 - \hat{p}_j)$, where $p_k$ is the probability that subject $k$ is in $T$. Assume that selection into $T$ or $C$ is exogenous, and the $p_k$ are well estimated.

We would like to know the average response if all study subjects were put into $T$. A sensible estimator is the weighted average response over the treatment group in the study,

$$\sum_{i\in T} Y_i w_i \Big/ \sum_{i\in T} w_i. \tag{8a}$$

Likewise, a sensible estimator for the average response if all subjects were put into $C$ is the weighted average over the study's control group,

$$\sum_{j\in C} Y_j w_j \Big/ \sum_{j\in C} w_j. \tag{8b}$$

These are approximations to the familiar Horvitz-Thompson estimators. The difference between (8a) and (8b) is a weighted contrast.

If selection is endogenous, or the weights are poorly estimated, the estimators in (8) are likely to be unsatisfactory. Even with exogenous selection, a large sample, and good estimates for the weights, variances may be large, and estimated variances may not be satisfactory—if there is a lot of variation in the weights across subjects. For instance, a relatively small number of subjects with large weights can easily determine the outcome, in which case the effective sample size is much reduced.

As a technical matter, the coefficient of the treatment variable in a weighted simple regression coincides with the weighted contrast (although the two procedures are likely to give different nominal variances). Anything distinctive about the weighted regression approach must involve the possibility of multiple regression when estimating the response equation. However, as we suggest above, it may be counterproductive to increase the analytic complexity by introducing multiple regression, variable selection, and the like.

## 9. CONTRASTS vs STRUCTURAL EQUATIONS

Linear causal models like (1) are called "response equations" or "structural equations." Implicitly or explicitly, the coefficients are often given causal interpretations. If you switch a subject from control to treatment, all else held constant, $X$ changes from 0 to 1. The response should then increase by the coefficient of $X$, namely, $b$. Similarly, if $Z_1$ is increased by one unit, all else held constant, the response should go up by $c_1$ units. In the papers by Robins and his school, the focus is quite different. Nothing is held constant. The objective is to estimate the average response—over all values of the confounders—if all subjects are put in treatment, or all subjects are put in control. When weights are used, it can take some effort to identify the estimands. For additional discussion of structural equations, see Freedman (2005).

## 10. CONCLUSIONS

Investigators who have a causal model that they believe in should probably just fit the equation to the data. If there are omitted variables but the propensity scores can be estimated with reasonable accuracy, weighting the regression should reduce bias.

On the other hand, weighting is likely to increase random error by a substantial amount, the nominal standard errors are often severely biased downward, and substantial bias can still be present in the estimated causal effects. Variation in the weights creates problems; the distribution of the weights should always be examined.

If the causal model is dubious but the selection model is believable, an option to consider is the weighted contrast between the treatment and control groups. However, this analysis may be fragile. Again, random errors can be large, and there can be serious problems in estimating the standard errors.

Going beyond continuous response variables and weighted least squares leads to additional complications. Each combination of response model and fitting procedure has to be considered on its own, to see what the weighted regression is going to estimate. Even with weighted least squares, some care is needed to identify estimands.

## REFERENCES

Arceneaux, K., Gerber, A. S., and Green, D. P. (2006). "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment," *Political Analysis*, 14, 37–62.

Bang, H. and Robins, J. M. (2005). "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–972.

Bluthenthal, R. N., Ridgeway, G., Schell, T., Anderson, R., Flynn, N. M. and Kral, A. H. (2006). "Examination of the Association between Syringe Exchange Program (SEP) Dispensation Policy and SEP Client-Level Syringe Coverage among Injection Drug Users," *Addiction*, 102, 638–646.

Crump, R. K., Hotz, V. J., Imbens, G. W., Mitnik, O. A. (2007). “Dealing with Limited Overlap in Estimation of Average Treatment Effects,” Technical report, Department of Economics, U. C. Berkeley.

Doss, H. and Sethuraman, J. (1989). “The Price of Bias Reduction When There Is No Unbiased Estimate,” *Annals of Statistics*, 17, 440–42.

Evans, S. N. and Stark, P. B. (2002). “Inverse Problems as Statistics.” *Inverse Problems* 18: R1–43.

Francesconi, M. and Nicoletti, C. (2006). “Intergenerational Mobility and Sample Selection in Short Panels,” *Journal of Applied Econometrics*, 21, 1265–1293.

Freedman, D. A. (2005). *Statistical Models: Theory and Practice*. New York: Cambridge University Press.

Freedman, D. A. (2006). “On The So-Called ‘Huber Sandwich Estimator’ and ‘Robust’ Standard Errors,” *The American Statistician*, 60, 299–302.

Freedman, D. A. (2008). “Some General Theory for Weighted Regressions,” http://www.stat.berkeley.edu/ census/wtheory.pdf

Glazerman, S., Levy, D. M., and Myers, D. (2003). “Nonexperimental versus Experimental Estimates of Earnings Impacts,” *Annals of the American Academy of Political and Social Science*, 589, 63–93.

Gozalo, P. L. and Miller, S. C. (2007). “Predictors of Mortality: Hospice Enrollment and Evaluation of Its Causal Effect on Hospitalization of Dying Nursing Home Patients,” *Health Services Research*, 42, 587–610.

Heckman, J. J. (1978). “Dummy Endogenous Variables in a Simultaneous Equation System.” *Econometrica* 46: 931–959.

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

Hernán, M. A., Brumback, B., Robins, J. M. (2001). “Marginal Structural Models to Estimate the Joint Causal Effects of Nonrandomized Treatments,” *Journal of the American Statistical Association*, 96, 440–448.

Hirano, K. and Imbens, G. W. (2001). “Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization,” *Health Services and Outcomes Research Methodology*, 2, 259–278.

Kang, J. D. Y. and Schafer, J. L. (2007). “Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data,” *Statistical Science*, 22, 523–39.

Leslie, S. and Theibaud, P. (2007). “Using Propensity Scores to Adjust for Treatment Selection Bias,” SAS Global Forum 2007: Statistics and Data Analysis, paper 184-2007.

Lunceford, J. K. and Davidian, M. (2004). "Stratification and Weighting Via The Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study,” *Statistics in Medicine*, 23, 2937–2960.

McNiel, D. E. and Binder, R. L. (2007). “Effectiveness of Mental Health Court in Reducing Recidivism and Violence,” *American Journal of Psychiatry*, 164, 1395–1403.

Peikes, D. N., Moreno, L., and Orzol, S. M. (2008). “Propensity Score Matching: A Note of Caution for Evaluators of Social Programs,” *The American Statistician*, 62, 222–31.

Ridgeway, G., McCaffrey, D., and Morral, A. (2006). “Toolkit for Weighting and Analysis of Nonequivalent Groups: A Tutorial for the TWANG Package,” RAND Corporation, Santa Monica.

Robins, J. M. and Rotnitzky, A. (1992). “Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers,” in N. Jewell, K. Dietz, V. Farewell (eds.), *AIDS Epidemiology—Methodological Issues*. Boston, MA: Birkhäuser, pp. 297–331.

Robins, J. M. and Rotnitzky, A. (1995). “Semiparametric Efficiency in Multivariate Regression Models with Missing Data,” *Journal of the American Statistical Association*, 90, 122–129.

Robins, J. M., Rotnitzky, A., Zhao, L. P. (1994). “Estimation of Regression Coefficients When Some Regressors Are Not Always Observed,” *Journal of the American Statistical Association*, 89, 846–866.

Robins, J. M., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). “Performance of Double-Robust Estimators When ‘Inverse Probability’ Weights Are Highly Variable,” *Statistical Science*, 22, 544–59.

Rotnitzky, A., Robins, J. M. and Scharfstein, D. O. (1998). “Semiparametric Regression for Repeated Outcomes with Nonignorable Nonresponse,” *Journal of the American Statistical Association*, 93, 1321–1339.

Sampson, R. J., Laub, J. H., and Wimer. C. (2006). “Does Marriage Reduce Crime? A Counterfactual Approach to Within-Individual Causal Effects,” *Criminology*, 44, 465–508.

Schonlau, M. (2006). “Charging Decisions in Death-Eligible Federal Cases (1995-2005): Arbitrariness, Capriciousness, and Regional Variation,” in S. P. Klein, R. A. Berk, and L. J. Hickman (eds.), *Race and the Decision to Seek the Death Penalty in Federal Cases*, Technical report #TR-389-NIJ, RAND Corportation, Santa Monica.

Tita, G. and Ridgeway, G. (2007). "The Impact of Gang Formation on Local Pattern of Crime," *Journal of Research on Crime and Delinquency*, 44, 208–237.

Wilde, E. T. and Hollister, R. (2007). "How Close Is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment." *Journal of Policy Analysis and Management*, 26, 455–77.

## The authors

David A. Freedman is Professor of Statistics, University of California, Berkeley, CA 94705-3860. freedman@stat.berkeley.edu

Richard A. Berk is Professor of Criminology and Statistics, University of Pennsylvania, Philadelphia, PA 19104-6286. berkr@sas.upenn.edu