Psychological Methods
2008, Vol. 13, No. 4, 279–313


1082-989X/08/$12.00 DOI: 10.1037/a0014268

# Average Causal Effects From Nonrandomized Studies: A Practical Guide and Simulated Example

Joseph L. Schafer
The Pennsylvania State University

Joseph Kang
Northwestern University

In a well-designed experiment, random assignment of participants to treatments makes causal inference straightforward. However, if participants are not randomized (as in observational study, quasi-experiment, or nonequivalent control-group designs), group comparisons may be biased by confounders that influence both the outcome and the alleged cause. Traditional analysis of covariance, which includes confounders as predictors in a regression model, often fails to eliminate this bias. In this article, the authors review Rubin's definition of an average causal effect (ACE) as the average difference between potential outcomes under different treatments. The authors distinguish an ACE and a regression coefficient. The authors review 9 strategies for estimating ACEs on the basis of regression, propensity scores, and doubly robust methods, providing formulas for standard errors not given elsewhere. To illustrate the methods, the authors simulate an observational study to assess the effects of dieting on emotional distress. Drawing repeated samples from a simulated population of adolescent girls, the authors assess each method in terms of bias, efficiency, and interval coverage. Throughout the article, the authors offer insights and practical guidance for researchers who attempt causal inference with observational data.

*Keywords:* nonequivalent control group design, propensity scores, Rubin's causal model



## Overview

The problem of causal inference in its simplest form can be described as follows. Consider a treatment that is either present or absent for each participant. The goal is to assess the average effect of the treatment on a subsequently measured outcome. We are likely to have a pretest (baseline) measure of the outcome and perhaps other variables that have been measured prior to treatment. We assume for simplicity that there are no missing values in the baseline measures, that there is no dropout prior to final outcome, that the treatments have been carried out as intended with full compliance, and that there is no interference among participants in the sense that the treatment received by one has no effect on the outcome of any other. Scientists widely agree that to establish a causal link, it is not sufficient to show a significant difference in average response for treated and untreated persons at the end of the study. One must also rule out the possibility that the discrepancy is due to systematic differences between the groups at baseline. If the treatment was randomly assigned as part of a designed experiment, that conclusion would be immediate, because randomization will, on average, eliminate those differences. If the assignment was beyond the researchers' control, however, the groups may not have been equivalent at the outset, and ruling out alternative explanations is challenging and controversial.

A popular strategy for ruling out alternatives is to measure as many confounders—pretreatment variables that may be related to both the treatment and the response—as possible and then estimate what the difference in average response between treated and untreated persons would be if the average values of the confounders in both groups were equal. This idea, which underlies classical analysis of covariance (ANCOVA) and regression, still prevails in many areas of social and behavioral science. Over the last 2 decades, however, another set of methods has taken hold and become widely accepted in economic analyses and health-outcomes research. Those methods are based on propensity scores (Rosenbaum & Rubin, 1983). ANCOVA,

Joseph L. Schafer, Department of Statistics and The Methodology Center, The Pennsylvania State University; Joseph Kang, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University.

This research was supported by National Institute on Drug Abuse Grant P50-DA10075. This research uses data from Add Health, designed by J. R. Udry, P. S. Bearman, and K. N. Harris and supported by National Institute of Child Health and Human Development Grant P01-HD31921.

Correspondence concerning this article should be addressed to Joseph L. Schafer, The Methodology Center, 204 East Calder Way, Suite 400, State College, PA 16801. E-mail: jls@stat.psu.edu

regression, and propensity scores share a common goal: to eliminate biases due to confounding. However, they attack the problem from different sides. ANCOVA and regression model relationships between the confounders and the outcome, whereas propensity scores model relationships between the confounders and treatment status.

To appreciate how these methods work, it is helpful to understand the connections between causal inference and missing data. Notions of causality pertain to how an intervention would have changed individuals' results. With two different treatments, we can imagine two possible responses for each participant. Causal effects may be defined as differences between these so-called potential outcomes (Rubin, 1974b). Because only one outcome can be observed for any participant, techniques for causal inference are, in essence, missing-data methods. Casting the problem in terms of missing data clarifies the goal of the analysis and helps us to gain insight from the missing-data literature.

In this article, we review techniques for estimating average causal effects (ACEs) from the standpoint of potential outcomes. Potential outcomes have been described by Gelman and Meng (2004), Holland (1986), Rosenbaum (2002), Rubin (2005), and others. Reviews from applied perspectives are given by D'Agostino (1998); Sobel (1995); Morgan and Winship (2007); West, Biesanz, and Pitts (2000); Winship and Morgan (1999); and Winship and Sobel (2004). Many of these articles emphasize methods based on propensity scores. In a series of publications over the last decade, James Robins and his colleagues have developed a theory of semiparametric estimation based on dual modeling of propensity scores and potential outcomes (Robins, Rotnitzky, & Zhao, 1994, 1995; Rotnitzky, Robins, & Scharfstein, 1998; van der Laan & Robins, 2003). These estimates have an interesting property known as double robustness, which means that the bias in an estimated causal effect still vanishes in a large sample if either of the two models (but not both) is wrong (Robins & Rotnitzky, 1995). Models for propensity scores and potential outcomes may be combined in various ways to achieve double robustness (Kang & Schafer, 2007). We review these methods and demonstrate how to use them to estimate a population ACE.

Throughout this article, we focus on the effect of a binary treatment on the mean of a numeric outcome. In practice, a treatment variable may be nominal, ordinal, or numeric, and the response could be any of these types as well; extensions to more general settings will be mentioned. We also make the simplifying assumption that *all confounders have been measured and are available to the analyst*. This assumption may approximately hold if an extensive set of measures known by subject-matter experts to be predictive of the treatment has been collected at the pretest. In other applications, this assumption will be highly questionable, for example, if only a few demographic variables are present. Even under the assumption of unconfoundedness, causal inference is not trivial; many solutions have been proposed, and there is no consensus among statisticians about which methods are best. By reviewing the available options, we hope to convey a basic understanding of how each method works, so that researchers can make informed choices and select techniques that are well suited to their applications. If nothing else, the best available answers under an assumption of no unmeasured confounders will establish useful benchmarks and points of departure for sensitivity analyses (Rosenbaum, 2002).

In the sections ahead, we review potential outcomes, define the ACE, and review the assumptions that must be made to estimate an ACE. We present a simulated case study that illustrates some of the difficulties of causal inference with observational data. We review ANCOVA and regression methods and explain the difference between a regression coefficient and a causal effect. We explain the role of propensity scores in causal inference, present nine different methods for estimating ACEs, and compare their performance in the simulated study. None of the estimates are difficult to compute, but formulas for their standard errors are tedious. These formulas, which are not readily available from other sources, are provided in the Appendix.

## The Potential-Outcomes Framework for Causal Inference

### *ACEs*

To characterize a treatment effect in a randomized experiment, Neyman (1923/1990) introduced multiple possible outcomes for each experimental unit. A similar notation was independently proposed by Rubin (1974b, 1978) in the context of observational studies, producing a framework that is often called Rubin's causal model (Holland, 1986). The idea has also been attributed to Haavelmo (1944). Other systems for causal inference have been proposed by Dawid (2000) and Pearl (2000), but Rubin's model is now the standard for causal thinking in statistics, epidemiology, and economics (Höfler, 2005; Pratt & Schlaifer, 1988; Rubin, 2005).

Let $T_i$ denote the treatment received by individual $i$. For simplicity, we suppose that there are only two conditions: treated ($T_i = 1$) and untreated ($T_i = 0$). Let $Y_{i1}$ denote the response if $T_i = 1$ and $Y_{i0}$ the response if $T_i = 0$. Covariates measured prior to treatment are denoted by $X_{i1}, X_{i2}, \ldots, X_{ip}$. We will often refer to the covariates as a column vector $X_i = (X_{i1}, \ldots, X_{ip})^T$, where the superscript $T$ denotes the transpose. Potential outcomes, treatment indicators, and covariates for a sample of $N$ individuals are shown in Table 1. If $T_i = 0$, then $Y_{i1}$ cannot be observed, and if $T_i = 1$, then $Y_{i0}$ cannot be observed. These missing potential outcomes have also been called counterfactuals (e.g., Greenland,

Table 1
*Rubin's Causal Model for Participants i = 1, . . ., N in a Hypothetical Observational Study: Covariates, Treatment indicators, Potential Outcomes, and Causal Effects*

| | Covariates | | | | Treatment | Potential outcomes | | Causal effect |
|---|---|---|---|---|---|---|---|---|
| $i$ | $X_{i1}$ | $X_{i2}$ | . . . | $X_{ip}$ | $T_i$ | $Y_{i0}$ | $Y_{i1}$ | $D_i = Y_{i1} - Y_{i0}$ |
| 1 | 5.0 | 0 | | 19.4 | 0 | 6.1 | 7.6[a] | 1.5[a] |
| 2 | 7.6 | 1 | | 21.5 | 1 | 7.2[a] | 7.9 | 0.7[a] |
| 3 | 4.8 | 1 | | 28.8 | 1 | 5.2[a] | 4.1 | −0.9[a] |
| 4 | 5.1 | 0 | | 12.7 | 0 | 4.8 | 7.1[a] | 2.3[a] |
| ⋮ | | | | | | | | |
| $N$ | 6.3 | 0 | | 17.0 | 0 | 6.9 | 8.3[a] | 1.4[a] |

[a] Denotes a value that is not observed.



Pearl, & Robins, 1999). The causal effect for an individual, defined as $D_i = Y_{i1} - Y_{i0}$, cannot be observed either; this is the "fundamental problem of causal inference" (Holland, 1986, p. 2). Nevertheless, we define the ACE in the population as the expected value of $D_i$,

$$ACE = E(D_i) = E(Y_{i1}) - E(Y_{i0}). \quad (1)$$

If all of the potential outcomes were seen, we would estimate the ACE by

$$\hat{ACE} = \frac{1}{N}\sum_{i=1}^{n} D_i = \frac{1}{N}\sum_{i=1}^{n} Y_{i1} - \frac{1}{N}\sum_{i=1}^{n} Y_{i0}. \quad (2)$$

The estimate in Equation 2 is not computable in a real study, but we regard it as a "gold standard" because it is unbiased and reasonably efficient regardless of how the data are distributed. Stronger assumptions are needed to estimate the ACE from data that are observed. If $T_i$ is independent of $Y_{i0}$ and $Y_{i1}$, as in a completely randomized experiment, then the treated and untreated groups are both representative samples from the full population. In that case, the difference between the mean response among the treated and the mean response among the untreated,

$$\hat{ACE} = \frac{\sum_i T_i Y_{i1}}{\sum_i T_i} - \frac{\sum_i (1 - T_i) Y_{i0}}{\sum_i (1 - T_i)}, \quad (3)$$

becomes an unbiased estimate of the ACE. [In Equation 3 and subsequent formulas, summations are taken over all participants $i = 1, \ldots, N$. Multiplying a summand by $T_i$ is equivalent to summing over those with $T_i = 1$, so that $\Sigma_i T_i Y_{i1}$ is the total of $Y_{i1}$ in the treated group, and $\Sigma_i T_i$ is the size of the treated group. Similarly, $\Sigma_i (1 - T_i) Y_{i1}$ is the total of $Y_{i0}$ in the control group, and $\Sigma_i (1 - T_i)$ is the size of the control group.]

In a typical observational study, however, it is unlikely that $T_i$ will be independent of $Y_{i0}$ and $Y_{i1}$. The treatments may have been selected by the individuals themselves, for reasons that are possibly related to the outcomes. With observational data, a good estimate of the ACE will make use of the covariates $X_i$ to help account for this dependence. Covariates do not redefine the causal effect that we are estimating. Rather, they help us solve the missing-data aspect of the problem. Covariates help us to recover information about the missing values, which in turn helps us to create an estimate that mimics the gold standard. Variables that are useful for a missing-data procedure but are not otherwise of interest are sometimes called auxiliary variables (Allison, 2001; Collins, Schafer, & Kam, 2001). When used wisely, auxiliary variables can lead to estimates with low bias and high efficiency.

In this framework, we do not assume that the causal effect for every individual is the same. Like any variable in the population, $D_i$ may vary and covary with other variables. An ACE near zero does not eliminate the possibility of seeing strong causal effects in subgroups because the treatment may be moderated by baseline variables. Conversely, a positive ACE does not preclude interactions that may lead to zero or negative effects in subgroups. Estimating an ACE is a good start, but richer analyses might attempt to describe $E(D_i|Z_i)$, where $Z_i$ is a set of individual characteristics that may include some of the $X_{ij}$s, and $E(\cdot \mid \cdot)$ denotes conditional expectation. The marginal structural model, introduced by Robins (1999), allows causal effects to vary in relation to other variables (Hernan, Brumback, & Robins, 2000). With a marginal structural model, the mean of each potential outcome may be modeled as logarithmic, logistic, or other nonlinear function of the covariates, as in a generalized linear model (Cohen, Cohen, West, & Aiken, 2003, chapter 13). This allows ACEs to be formulated as multiplicative effects, odds ratios, and so forth—which may be useful when the response variable is discrete. If a causal effect is seen to vary across groups, however, this should not be taken as evidence that the characteristics defining the groups *caused* the difference. Causal statements about $Z_i$

would require an expanded notation that defines potential outcomes for distinct values of $Z_i$.

### *ACE for the Treated*

The population ACE is the average difference in response between the following two scenarios: First, everyone in the population receives $T_i = 1$; second, everyone in the population receives $T_i = 0$. If the treatment $T_i = 1$ may conceivably be applied to the whole population—as in a universal intervention designed to prevent problem behavior in school children—then this average difference will be meaningful. If the treatment is inapplicable or irrelevant to large segments of the population, however, then alternatives to the ACE are worth considering. One useful alternative is the ACE among the treated,

$$ACE_1 = E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 1),$$

which measures how the treatment affects those who are receiving it (Hirano & Imbens, 2001; Winship & Sobel, 2004). Economists have argued that $ACE_1$ is often more relevant than ACE for discussing implications of policy (Heckman, Smith, & Clements, 1997). Suppose, for example, that we want to assess the effect of workplace smoking on employees' health care expenditures. If the study is meant to inform us about the possible benefits of a ban on workplace smoking, then the population of interest would be those who currently smoke in the workplace, because the ban would presumably have little impact on those who do not engage in that behavior. A ban would switch all values of $T_i = 1$ to $T_i = 0$, affecting only those who currently have $T_i = 1$. Similarly, one can also define an ACE for the untreated,

$$ACE_0 = E(Y_{i1} \mid T_i = 0) - E(Y_{i0} \mid T_i = 0),$$

which describes the impact of switching everyone with $T_i = 0$ to $T_i = 1$. The gold-standard estimate of $ACE_1$, which we would use if all values of $Y_{i0}$ were seen, is

$$\hat{ACE}_1 = \frac{\sum_i T_i (Y_{i1} - Y_{i0})}{\sum_i T_i}. \quad (4)$$

### *Assumptions Needed to Estimate an ACE*

To estimate an ACE from observable quantities, we need to make several assumptions. First, we must assume that the treatment applied to any individual does not affect the outcome of any other individual. This is called the stable unit treatment value assumption (Rubin, 1980). This assumption may be violated in multilevel settings in which participants interact in close proximity (e.g., in classrooms or schools), and the treatment received by one person may impact those around him or her. Hong and Raudenbush (2005) extended stable unit treatment value assumption to account for possible interference among participants in a multilevel study, but we do not consider that extension here.

Any method for estimating an ACE will also require assumptions about the relationships between $T_i$ and the potential outcomes. Strict independence between $T_i$ and $(Y_{i0}, Y_{i1})$, as in a completely randomized experiment, is untenable in most observational studies, so the assumption is weakened in the following way. A treatment mechanism is said to be *unconfounded* given a set of covariates $X_i$ if, conditioning upon $X_i$, the potential outcomes $Y_{i0}$ and $Y_{i1}$ are jointly independent of $T_i$ (Rubin, 1978). Unconfoundedness, which is also called strong ignorability (Rosenbaum & Rubin, 1983), means that any relationship between the potential outcomes and the treatment status is fully explained by $X_i$, that is, that there are no unmeasured confounders. It also implies that the unseen potential outcomes are missing at random (MAR; Rubin, 1976), an assumption that is often made for missing-data problems. Implications of MAR are discussed by Schafer and Graham (2002). One important point from that discussion is that MAR becomes more plausible as the number of covariates in $X_i$ grows. Introducing more covariates reduces the residual dependence of $T_i$ on $Y_{i0}$ and $Y_{i1}$, which helps to eliminate the selection bias arising from nonrandom treatment assignment.

Unconfoundedness, like MAR, cannot be verified or contradicted on the basis of observed data (Little & Rubin, 2002). Its reasonableness can only be evaluated by expert judgment of how individuals came to receive their treatments. If unconfoundedness is violated, the missing potential outcomes will be missing not at random. In that case, estimating an ACE requires a model in which $T_i$ depends on the potential outcomes. Fitting these models is not trivial, and other unverifiable assumptions must inevitably be made. For discussion on missing values that are missing not at random, see Little and Rubin (2002) and Schafer and Graham (2002). Large sections of the textbook by Rosenbaum (2002) are devoted to sensitivity analyses for investigating the impact of departures from unconfoundedness. In one popular technique, the analyst posits the existence of an omitted variable that jointly affects the treatment and the potential outcomes, and then observes how strongly the variable must be related to the treatment to make the ACE disappear (Cornfield et al., 1959).

Unconfoundedess may also be violated if some of the variables in $X_i$ were measured after the treatment and were influenced by the treatment. In our framework, we assume that each person's covariates would have been the same whether that person had $T_i = 0$ or $T_i = 1$. In general, one should not adjust for variables measured after the treatment as if they were confounders, because doing so could introduce posttreatment selection bias (Rosenbaum, 1984). Studies in which $X_i$ and $T_i$ are collected simultaneously may also

be problematic. If the covariates include self-reported measures of psychological states, then—even if the wording of the items refers to a period of time before the treatment—participants' recollections of those states may be influenced by the treatment. Measuring $T_i$ at a later time does not always solve this problem, because, depending on the context, the decisive chain of events that led to $T_i = 0$ or $T_i = 1$ may have occurred before some of the $X_{ij}$s were determined.

Unconfoundedness is helpful, but it is not enough. To estimate an ACE, we need to model the distribution of $T_i$ given $X_i$, the distributions of $Y_{i0}$ and $Y_{i1}$ given $X_i$, or both. Misspecification of the form of these models can lead to bias in the resulting ACE estimates, and measurement errors in covariates may act as omitted variables.

From the pattern of missing values in Table 1, it is apparent that the relationship between the potential outcomes—more precisely, the partial correlation between $Y_{i0}$ and $Y_{i1}$ given the covariates—cannot be estimated from the observed data (Rubin, 1974a). Without introducing prior information about this relationship, a joint distribution for $Y_{i0}$ and $Y_{i1}$ cannot be identified. Because the ACE is defined in terms of the marginal means of $Y_{i0}$ and $Y_{i1}$, however, joint modeling is not necessary. Steyer (2005) proposed an extension of the potential-outcomes framework that uses repeated measurements on individuals to glean more information about the unseen responses, permitting the construction of a full joint model. We could, for example, use the baseline measurement for an individual who later received the treatment $T_i = 1$ as a proxy for $Y_{i0}$. If multiple waves of pretest measurements are available, we may use the pretest trajectory to forecast what a treated person's outcomes might have been in the absence of treatment. Doing so requires strong assumptions. In particular, it requires us to correctly specify the form of the trajectories, and it requires this form to remain constant over the study period so that extrapolation to the posttreatment occasions is valid. Models of this type are reviewed by Winship and Morgan (1999). Haviland, Nagin, and Rosenbaum (2007) used pretreatment trajectories in the outcome variable to define substantively meaningful groups, and then they performed propensity-based analyses to estimate ACEs within these groups.

Finally, to estimate an ACE, we will have to assume that each person in the population could have received either treatment (Rubin, 1978). If the probability of $T_i = 0$ or $T_i = 1$ is zero for some individuals, then it may not be meaningful to speak of causal effects for them, because they could not have experienced the alternative treatment.

## A Simulated Observational Study

### *Dieting and Emotional Distress*

To illustrate and evaluate competing methods for estimating ACEs, we simulated a study of the psychological effects of dieting, a behavior that would be difficult or impossible to assign in a randomized controlled experiment. Dieting—defined as an intentional, temporary reduction in caloric intake—is generally regarded as an ineffective long-term strategy for weight loss (Hill, 2004; Katz, 2005). Moreover, it has been associated with negative psychological outcomes, including depression and anxiety (Kovacs, Obrosky, & Sherrill, 2003; Warren & Cooper, 1988); decreased cognitive performance (Green & Rogers, 1995); and onset of binge eating, anorexia nervosa, and bulimia nervosa (Hsu, 1996; Patton, Selzer, Coffey, Carlin, & Wolfe, 1999; Wilson, 1993). Cross-sectional studies have consistently shown that adolescent girls who diet have higher levels of negative affect and psychological distress than those who do not (Neumark-Sztainer & Hannan, 2000). Heatherton and Polivy (1992) hypothesized that dietary restraint and failed attempts at weight control lead to distress. The causal link, however, has not been supported by longitudinal data. Rosen, Tacy, and Howell (1990) found that dieting prospectively predicted increased stress but not general psychological distress. Johnson and Wardle (2005) found that dietary restraint had no significant effect on distress 1 year later after controlling for measures of body dissatisfaction. Those researchers relied on multiple regression and reported results in terms of regression coefficients rather than ACEs.

### *A Simulated Study Based on Add Health*

Our simulated study loosely resembles the National Longitudinal Study of Adolescent Health (Add Health)—a representative sample of American middle and high school students measuring a broad array of health characteristics, behaviors, and attitudes (Udry, 2003). Add Health has complications that make analysis challenging, including multilevel structure (students within schools), unequal probabilities of selection at multiple stages, missing items, and dropout between waves. Methods for dealing with these complications are available (e.g., Zanutto, 2006), but applying them all within this article would detract from the issues we want to address. Rather than using Add Health directly, we devised a simulation that mimics some of its variables but not its complex design, nonresponse, or dropout.

Using probability distributions estimated from Add Health, we generated random data to create a synthetic population of 1 million adolescent girls. From this population, we drew simple random samples of $N = 6{,}000$ girls with no missing values or dropout and applied estimation procedures to each sample. Our choice of $N = 6{,}000$ was motivated by the fact that Add Health has usable diet-related data at Waves I and II for about 6,500 girls. Repeating the sampling procedures many times, we were able to see how various methods for estimating ACEs reproduced the known effects in the population.

Because no Add Health participants appear in this simu-

lation, we are able to provide interested readers with a sample of our data. Our first random sample of 6, 000 girls, and the programs we used (written in the R language) to perform the various analyses, have been placed on a website at http://www.stat.psu.edu/~jls/causal/index.html.

Sample sizes in the thousands are not uncommon in epidemiologic and public health contexts, but research in psychology typically involves fewer participants. Jaccard and Wan (1996) have reported a median *N* among studies in psychology involving regression analyses of approximately 200. The large sample size in our simulation affects our results in two ways. First, it heightens the impact of bias when models are misspecified or other assumptions (e.g., unconfoundedness) are violated. A useful rule-of-thumb is that bias begins to seriously impair the performance of confidence intervals and significance tests when the magnitude of the bias in an estimate exceeds 40%–50% of its standard error (Collins et al., 2001). In the sections ahead, the differences in performance among competing methods for estimating ACEs that we report are mainly due to varying degrees of bias. Indeed, reduction of bias is often cited as the primary concern in causal inference from nonrandomized studies (see, e.g., the essay by Rubin, 2006, pp. 460–462), so the emphasis on bias in this simulation is warranted. For studies with smaller *N*, however, researchers should be aware that the differences in bias among methods may be less dramatic than what we report, and considerations of efficiency (variance reduction) and power may deserve greater attention. Second, our large *N* guarantees that the aymptotic (large-sample) approximations that we use to compute standard errors and confidence intervals are appropriate. Smaller values of *N* may impair the performance of these approximations, again damaging the performance of intervals and tests. Moreover, with smaller *N*, researchers may need to be more sparing in choosing baseline covariates from the pool of potential confounders to include in their adjustment procedures. The special problems that arise in causal inference with nonrandomized studies of small to moderate size have received relatively little attention in the literature to date and are an important topic for future study.

## *Variables*

Variables in the synthetic population resemble items from Add Health Waves I (1994–1995) and II (1995–1996). Dieting is measured by an item at Wave I asking whether the participant dieted to maintain or lose weight in the last 7 days. By this criterion, 20% of the girls were dieting. The wording of this item is not ideal, because the 7 days are prior to the measurement of the baseline variables. Covariates measured after the treatment should not be regarded as confounders; doing so may introduce posttreatment selection bias (Rosenbaum, 1984). In our simulation, we prevented this bias by generating values for all baseline variables first and then generating values for the dieting variable given the baseline variables.

The outcome is a composite measure of emotional distress at Wave II based on a 19-item feelings scale. Participants were asked whether they felt bothered, tired, depressed, lonely, and so forth with response categories ranging from 0 (*never or rarely*) to 3 (*most or all of the time*). Items whose wording was positive (e.g., happy or hopeful about the future) were reverse-scored to make higher values consistent with greater distress. Using a procedure similar to that of Resnick et al. (1997), we averaged the items to form a single measure whose values are continuously distributed between 0 and 3. Internal reliability of the measure was high (Cronbach's $\alpha = .88$). The feelings scale was also administered at baseline (Wave I), and the correlation between the measures across the waves is strong ($r = .60$). The baseline measure of emotional distress plays a crucial role in most of our methods for estimating ACEs, because it is highly predictive of the outcome and also related to dieting.

Under Rubin's causal model, each girl has two potential values of the outcome at Wave II: one if she diets, another if she does not. For each girl in the population, we simulated two versions of the outcome corresponding to these two conditions. The simulated values of these variables for the first five girls in the population are shown in Table 2. In this table, $Y_{i0}$ denotes emotional distress after not dieting, $Y_{i1}$ denotes emotional distress after dieting, and $T_i$ denotes the dieting behaviors (0 = no, 1 = yes). The distress actually seen at Wave II is $Y_i$, and the causal effect of dieting is $D_i = Y_{i1} - Y_{i0}$. Notice that the first two girls dieted, and it slightly increased their emotional distress ($D_i = 0.02$ and 0.01). The third girl also dieted, but her dieting was beneficial ($D_i = -0.10$). The fourth girl did not diet, and wisely so, because if she had, her distress would have increased ($D_i = 0.26$).

Table 2
*Simulated Binary Causal Variable and Potential Outcomes for the First Five Individuals (i = 1, . . ., 5) in the Artificial Population of One Million Adolescent Girls*

| Individual *i* | $Y_{i0}$ | $Y_{i1}$ | $T_i$ | $Y_i$ | $D_i$ |
|---|---|---|---|---|---|
| 1 | 1.27 | 1.29 | 1 | 1.29 | 0.02 |
| 2 | 0.27 | 0.28 | 1 | 0.28 | 0.01 |
| 3 | 0.17 | 0.07 | 1 | 0.07 | −0.10 |
| 4 | 2.18 | 2.44 | 0 | 2.18 | 0.26 |
| 5 | 1.05 | 0.97 | 0 | 1.05 | −0.08 |
| *M* in population | 0.66 | 0.66 | 0.20 | 0.65 | 0.00 |
| *SD* in population | 0.47 | 0.49 | 0.40 | 0.47 | 0.16 |

*Note.* $Y_{i0}$ = emotional distress at Wave II without dieting; $Y_{i1}$ = emotional distress at Wave II with dieting; $T_I$ = dieting behavior (0 denotes no dieting; 1 denotes dieting); $Y_I$ = emotional distress seen at Wave II; $D_I$ = causal effect. Also shown are the means and standard deviations in the population.

The fifth girl did not diet, but if she had, her distress would have decreased ($D_i = -0.08$).

We also created a set of confounders. One confounder is the baseline measure of emotional distress. Other known correlates of dieting include age, race, ethnicity, self-perceived weight, body image (Gerner & Wilson, 2005), and self-esteem (Neumark-Sztainer & Hannan, 2000). Thirteen variables related to these concepts that resemble variables from Add Health at Wave I were simulated for each girl in the population. Descriptions of these confounders and their population means and standard deviations are shown in Table 3.

### *How the Population Data Were Generated*

When creating the population, we avoided simple parametric models commonly used by data analysts, because naturally occurring populations rarely conform to such assumptions. Special efforts were made to capture nonnormal shapes, nonlinearities, and interactions. These complexities, however, were not reflected in the analytic procedures applied to the samples. The methods used to estimate ACEs were based on simpler models that did not precisely agree with those used to create the population, which adds a degree of realism.

Distributions for the synthetic population were estimated from 6,503 girls in the Add Health grand sample who participated in the home interview at Waves I and II. First, we randomly sampled 1 million values of grade with five levels (7, . . ., 11) and race/ethnicity with three levels (Black, non-Black Hispanic, other) using the unweighted sample proportions from the $5 \times 3$ contingency table. Next, we drew baseline measures of emotional distress by random sampling with replacement within the 15 cells defined by grade and race/ethnicity. We then simulated each of the remaining confounders listed in Table 2 (self-rating of overall health, self-rating of weight, etc.) using a propotional-odds logistic model (Agresti, 2002) that included main effects for the previous confounders in the sequence and all two-way interactions among them. The 13 confounders in our population closely resemble their Add Health counterparts with respect to marginal distributions, pairwise relationships, and three-way associations.

Given these 13 confounders, we generated the potential outcomes $Y_{i0}$ and $Y_{i1}$ by rich regression models—one for dieters and one for nondieters—fit to a transformed version of emotional distress at Wave II. The transformation ensured that the predicted values remained between 0 and 3. Each model included main effects for all confounders and many interactions among them. Effects of emotional distress at baseline were described by piecewise linear trends with changes in slope at the 20th, 40th, 60$^{th}$, and 80th percentiles of the baseline values. Each 5-point Likert scale item was described by a set of four dummy indicators, allowing for trends of arbitrary shape. A generous subset of main effects and two-way interactions was selected for each regression model by an automatic stepwise search. After fitting the models to Add Health, we applied them to the simulated population to compute predicted values of $Y_{i0}$ and $Y_{i1}$ for each girl. To the predicted values, we added correlated random residuals from a nonnormal bivariate distribution whose marginals exactly matched the empirical distributions of the residuals in Add Health. The simulated outcomes were then transformed back to the original measurement scale to produce $Y_{i0}$ and $Y_{i1}$.

Table 3
*Confounding Variables in Artificial Population of One Million Adolescent Girls, With Descriptions, Means, and Standard Deviations*

| Name | Description | *M* | *SD* |
|---|---|---|---|
| DISTR.1 | Emotional distress at Wave I (minimum = 0, maximum = 2.84) | 0.64 | 0.43 |
| BLACK | 1 = Black, 0 = otherwise | 0.23 | 0.42 |
| NBHISP | 1 = non-Black Hispanic, 0 = otherwise | 0.16 | 0.36 |
| GRADE | Grade in school at Wave I (7, . . ., 11) | 9.20 | 1.39 |
| SLFHLTH | Self-rating of overall health (1 = *excellent*, 2 = *very good*, 3 = *good*, 4 = *fair*, 5 = *poor*) | 2.19 | 0.92 |
| SLFWGHT | Self-rating of weight (1 = *very underweight*, 2 = *slightly under*, 3 = *about right*, 4 = *slightly over*, 5 = *very over*) | 3.33 | 0.79 |
| WORKHARD | "When you get what you want, it's usually because you worked hard for it" (1 = *strongly agree*, . . ., 5 = *strongly disagree*) | 2.15 | 0.91 |
| GOODQUAL | "You have lots of good qualities" (1 = *strongly agree*, . . ., 5 = *strongly disagree*) | 1.81 | 0.69 |
| PHYSFIT | "You are physically fit" (1 = *strongly agree*, . . ., 5 = *strongly disagree*) | 2.29 | 0.94 |
| PROUD | "You have a lot to be proud of" (1 = *strongly agree*, . . ., 5 = *strongly disagree*) | 1.78 | 0.77 |
| LIKESLF | "You like yourself just the way you are" (1 = *strongly agree*, . . ., 5 = *strongly disagree*) | 2.19 | 1.02 |
| ACCEPTED | "You feel socially accepted" (1 = *strongly agree*, . . ., 5 = *strongly disagree*) | 2.18 | 1.02 |
| FEELLOVD | "You feel loved and wanted" (1 = *strongly agree*, . . ., 5 = *strongly disagree*) | 1.82 | 0.85 |

Finally, we created the treatment indicator by fitting a rich binary regression model to predict dieting from the confounders. Our model used a complementary log–log link function (Agresti, 2002) and included main effects for the confounders, a cubic trend for baseline distress, and numerous interactions. We deliberately chose the complementary log–log link to ensure that all propensity models we subsequently fit to the data, which use the logistic link, would be misspecified. Applying this model to the population, we computed the probability of dieting for each girl from her covariates and generated $T_i$ by the flip of a weighted coin. This treatment mechanism did not involve the potential outcomes, and thus it is unconfounded. This may be the most unrealistic aspect of our simulation. Many other psychosocial variables known to influence girls' decisions to diet—variables that describe family and peer criticism of weight (Levine, Smolak, & Hayden, 1994) and thin-ideal internalization (Stice, Shaw, & Nemeroff, 1998), for example—were not collected in Add Health. To the extent that these omitted variables were related to the treatment and to the potential outcomes, unconfoundedness would not hold in a real version of this study.

### *ACEs in the Synthetic Population*

The average effect of dieting on emotional distress in our synthetic population is ACE = 0.003. Compared with the standard deviation of the observed distress score at Wave II ($SD = 0.47$), the effect is practically insignificant. However, this does not mean that dieting had no impact. Dieting increased distress for some and decreased it for others. A quantity that seems more relevant from a standpoint of health policy is the ACE among the girls who actually dieted. If dieting is a behavior to be discouraged, an intervention that successfully eliminates it would produce a mean change of $ACE_1$ within this group. The average effect of dieting among dieters in our population is $ACE_1 = -0.022$, which means that, on average, it led to a slight decrease in emotional distress for the girls who were doing it. If we examine the effects in other subsets of the population, we find some that are much larger. For example, the average effect of dieting among girls who responded *strongly disagree* to the item "You have lots of good qualities" (GOODQUAL = 5) is 0.210. Dieting interacted with this covariate to increase distress among girls with strongly negative self-image.

We do not claim that these results hold for girls in the U.S. population. Similar effects might be found there, because our synthetic data are based on a representative sample. Although we do not present these findings as substantively meaningful, they underscore the fact that an ACE can be a crude summary of causal effects in a large and diverse population.

## Mean Comparisons, ANCOVA, and Regression

### *Method 1: Simple Difference in Means*

We now examine methods for estimating ACEs and evaluate their performance in our simulated observational study. To facilitate comparisons among the methods, we present two tables near the end of this article—Table 6, which reports estimates and standard errors from the first sample of $N = 6{,}000$, and Table 7, which summarizes performance over 5,000 samples. Readers are encouraged to refer to these tables as they peruse the remaining sections.

The first quantity that we examine is the simple difference between the mean distress scores observed for the dieters and the nondieters (Equation 3). This simple difference is an unbiased estimate of the *prima facie* effect, defined as

$$PFE = E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 0) \tag{5}$$

(Holland, 1986). The *prima facie* effect becomes equal to the true ACE when the treatment is randomly assigned. More generally, Winship and Morgan (1999) have shown that the difference between the two effects can be written as

$$PFE - ACE = E(Y_{i0} \mid T_i = 1) - E(Y_{i0} \mid T_i = 0) + (1 - \pi)(ACE_1 - ACE_0),$$

where $\pi = P(T_i = 1)$ is the proportion of treated persons in the population. The bias in the simple difference in means as an estimate of the ACE can thus be ascribed to two sources: the difference in mean response between the two population groups in the absence of treatment, and the difference between the group-specific ACEs. With randomization, treated and untreated persons are drawn from the same population, and both of these types of bias vanish.

Our first sample of $N = 6{,}000$ girls yielded 1,220 dieters and 4,780 nondieters. Box plots of the emotional distress at Wave II for these two groups are shown in Figure 1a. The average for dieters (0.703) exceeds the average for nondieters (0.644), so girls who dieted tended to experience greater distress. The pooled standard error for this difference is $SE = 0.015$, and the usual $t$-test is significant, $t(5, 998) = 3.99$, $p < .001$. The difference between the means, 0.060, would unbiasedly estimate the ACE if this was a randomized experiment. In this nonrandomized observational study, however, girls who chose to diet and girls who did not were different a priori. For example, the average difference in emotional distress at Wave I for the two groups is 0.092, which is also highly significant, $t(5, 998) = 6.82$, $p < .001$. Differences between the dieters and nondieters in their baseline measures underscore the need to perform adjustments to mitigate bias due to confounding.

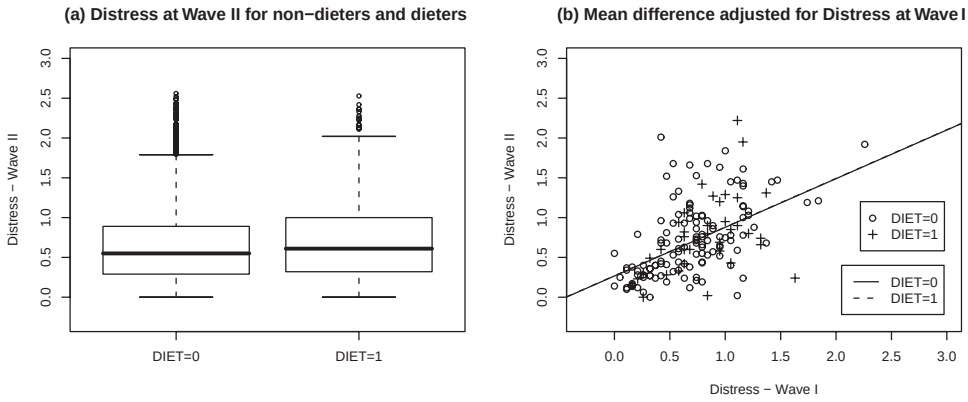


*Figure 1.* (a) Box plots of emotional distress at Wave II for nondieters (DIET = 0) and dieters (DIET = 1) in a sample of $N = 6{,}000$ girls from the synthetic population. The center lines of each box represent the sample median; the edges of the box represent the 25th and 75th percentiles; and circles denote outliers, defined as observations that lie more than 1.5 times the interquartile range above the 75th percentile. (b) Parallel regression lines from a linear analysis of covariance model adjusting for emotional distress at Wave I. (To avoid overplotting, only 150 randomly selected points are shown.)

### *Method 2: Regression and ANCOVA*

Regression adjustments, also known as the ANCOVA, use potential confounders as predictors in a regression model (Cook & Campbell, 1979). The most common version of ANCOVA assumes that relationships between the outcome and the potential confounders are linear. It also assumes that the slopes for the confounders among treated and untreated persons are identical. Let $Y_i$ denote the response and $T_i$ the treatment for participant $i$. Let $X_{i1}$, $X_{i2}$, . . ., $X_{ip}$ represent the participant $i$'s scores on a set of $p$ potential confounders collected at baseline. The simple linear version of ANCOVA assumes that

$$Y_i = \alpha + \theta T_i + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_p X_{ip} + \epsilon_i,$$

where $\varepsilon_i$ is a residual error with mean zero and constant variance. Let us collect the $X$s and $\beta$s into column vectors, $X_i = (X_{i1}, \ldots, X_{ip})^T$ and $\beta = (\beta_1, \ldots, \beta_p)^T$, where superscript $T$ indicates transpose (i.e., changing a row into a column). The model can now be written as

$$E(Y_i \mid T_i, X_i) = \alpha + T_i\theta + X_i^T\beta. \quad (6)$$

The coefficients $\alpha$, $\theta$, and $\beta$ are typically estimated by ordinary least squares (OLS). The treatment effect, $\theta$, is the average effect of switching from $T_i = 0$ to $T_i = 1$, adjusted for $X_i$. This model is so popular that it is sometimes assumed to be the only type of ANCOVA. In modern usage, however, ANCOVA refers to a general class of regression models that may include interactions, nonlinear relationships, and heteroscedastic errors (Huitema, 1980; Little, An, Johanns, & Giordani, 2000).

We first applied the model of Equation 6 to our sample of $N = 6{,}000$, controlling only for the emotional distress score recorded at baseline ($p = 1$). With this adjustment, the estimated effect of dieting on distress at Wave II dropped from .060 (the simple mean difference) to .003, and the standard error dropped from .015 to $SE = .012$. This ANCOVA model, as shown in Figure 1b, describes the relationship between the response and baseline measurements as two parallel lines, and the vertical distance between the lines is the estimated treatment effect. The effect is so small that the two lines are indistinguishable. With two or more covariates, the regression lines in Figure 1b become parallel planes. Including all $p = 13$ covariates listed in Table 3, the estimated effect becomes $-.014$ ($SE = .013$).

### *Conceptual Difficulties With ANCOVA for Observational Studies*

ANCOVA was developed by R. A. Fisher during the 1930s for randomized experiments. In those settings, differences between the groups at baseline are not systematic and vanish on average over repeated randomizations. In a randomized experiment, ANCOVA increases the precision of estimated treatment effects. When applied to data from a nonrandomized study, however, the method assumes a different purpose: to reduce bias due to preexisting systematic differences between the groups (Cochran, 1983; Cook & Campbell, 1979; Rubin, 1973, 1979). ANCOVA is just one of many possible techniques for bias reduction in observational studies. Its validity rests upon the assumption that the mean response varies linearly with $X_i$ with identical slopes when $T_i = 0$ and $T_i = 1$. The performance of ANCOVA can be sensitive to departures from these assumptions, particu-

larly if the distributions of confounders in the two groups are very different (Rubin, 1984, 1990).

From our perspective, the main problem with ANCOVA is that its connection to causality becomes muddled *when the model is wrong*. The ANCOVA treatment effect,

$$\theta = E(Y_i | T_i = 1, X_i) - E(Y_i | T_i = 0, X_i), \qquad (7)$$

is the difference in average response between treated and untreated persons with identical values of $X_i$. This interpretation of θ holds if Equation 6 accurately depicts the dependence of $Y_i$ on $T_i$ and $X_i$ in the population, that is, if the relationships between $Y_i$ and the covariates in $X_i$ are linear, and if the difference in mean response between $T_i = 0$ and $T_i = 1$ does not vary with the covariates. In real examples, no treatment is equally effective for all types of individuals, and the linear version of ANCOVA shown in Equation 6 will at best only roughly approximate the true mean structure. Moreover, the meaning of θ changes when variables are added or removed. ANCOVA posits a model for $E(Y_i|T_i,X_i)$, and this model depends heavily on which covariates included in $X_i$ and how the effects of these covariates are parameterized. In contrast, the notion of an ACE—what would happen if everyone received the treatment versus if they did not—is a simple, intuitive idea that is well-defined apart from any covariates or parametric models.

The ANCOVA treatment effect is an average difference in response *between two different groups of individuals*, adjusting for differences in their covariates. Causal inferences, on the other hand, are about changes in response when different treatments are applied *to the same individuals* (Rubin, 2004). The two concepts are related and, under special conditions, produce identical values for a population causal effect, but in general they are not the same.

## *A Geometric Interpretation of the ACE*

We have described the treatment effect in linear ANCOVA as a vertical distance between parallel regression lines or planes. A vertical distance between lines or planes can also represent an ACE, but the ACE is subtly different. ANCOVA conditions on the covariates in the model, but the ACE averages over the covariates.

To visualize the ACE, let us momentarily suppose that the potential outcomes are related to the covariates by linear regression models. Suppose that $E(Y_{i0}|X_i) = \alpha_0 + X_i^T\beta_0$ and $E(Y_{i1}|X_i) = \alpha_1 + X_i^T\beta_1$, where $\alpha_0$ and $\alpha_1$ are the intercepts, and $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})^T$ and $\beta_1 = (\beta_{11}, \ldots, \beta_{1p})^T$ are now vectors of slopes specific to the $T_i = 0$ and $T_i = 1$ groups. Although $Y_{i0}$ and $Y_{i1}$ are different variables, their units of measurement are the same, so we may plot them on the same axis. Regression planes for $Y_{i0}$ and $Y_{i1}$ with $p = 2$ covariates, $X_i = (X_{i1}, X_{i2})^T$, are shown in Figure 2.

In this picture, the distributions of the covariates when

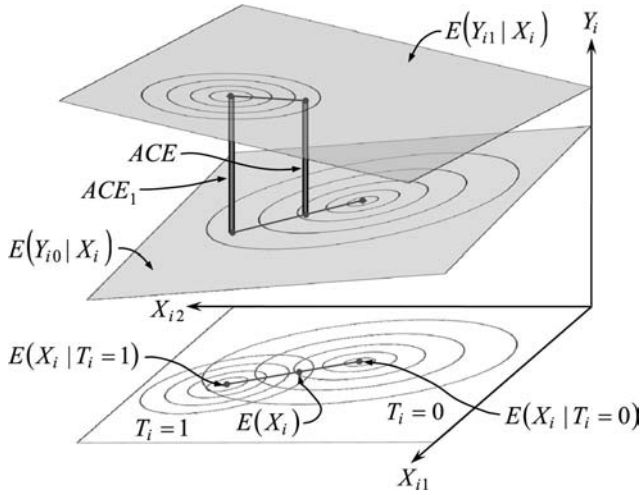


*Figure 2.* Nonparallel regression planes for a pair of potential outcomes, $Y_{i0}$ and $Y_{i1}$, given a pair of covariates $X_i = (X_{i1}, X_{i2})$. Joint distributions of the covariates among treated ($T_i = 1$) and untreated ($T_i = 0$) individuals are represented by concentric ellipses; $E(X_i|T_i = 1)$ and $E(X_i|T_i = 0)$ represent the average values of the covariates in the treated and untreated populations, respectively; and $E(X_i)$ represents the average value of the covariates in the population. The average causal effect (ACE) is represented by the vertical distance between the planes at $X_i = E(X_i)$. The ACE among the treated, $ACE_1$, is represented by the vertical distance between the planes at $X_i = E(X_i|T_i = 1)$. The ACE among the untreated, $ACE_1$, cannot be estimated in this example.

$T_i = 0$ and $T_i = 1$ are represented by elliptical contours that are analogous to lines of constant elevation on a topographical map. The highest concentrations of individuals are found at the centers of the ellipses; as we move out from the centers, the densities drop, and observations become more sparse. Elliptical contours are consistent with bivariate normal distributions, but we are not suggesting that the covariates must be normal; the interpretations that follow are valid regardless of how the covariates are distributed. In a randomized experiment, the distributions of covariates for $T_i = 0$ and $T_i = 1$ would coincide, because $T_i$ would be independent of $X_i$. In an observational study, however, $T_i$ is usually related to the covariates, so we have represented these distributions as different but partly overlapping. The average value of the covariates among treated persons is $E(X_i|T_i = 1)$, and the average among untreated persons is $E(X_i|T_i = 0)$. The overall average, $E(X_i)$, lies along the line segment connecting $E(X_i|T_i = 0)$ and $E(X_i|T_i = 1)$ with its position depending on the relative sizes of the two groups. Notice that in this drawing, $E(X_i|T_i = 1)$ and $E(X_i)$ are covered by both covariate distributions, but $E(X_i|T_i = 0)$ lies outside the distribution for $T_i = 1$. The degree of overlap has crucial implications for causal inference, as we soon discuss.

For any two random variables *A* and *B*, the identity

$E(A) = E(E(A|B))$ holds provided that these expectations exist. From this identity, it follows that

$$ACE = E(Y_{i1}|X_i = E(X_i)) - E(Y_{i0}|X_i = E(X_i)) = (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)^T E(X_i). \quad (8)$$

The ACE for the entire population is the vertical distance between the regression planes $E(Y_{i1}|X_i)$ and $E(Y_{i0}|X_i)$ above the point $E(X_i)$. The ACE among the treated, $ACE_1$, is represented by the vertical distance between the planes above the point $E(X_i|T_i = 1)$. If the regression planes happen to be parallel—that is, if each element of $\beta_0$ equals the corresponding element of $\beta_1$—then the vertical distance between the two planes will be constant $(\alpha_1 - \alpha_0)$, and $ACE_1$ will equal ACE. Under our linearity assumptions, the ACE among the untreated, $ACE_0$, is represented by the vertical distance between the planes above the point $E(X_i|T_i = 0)$. This quantity is not shown in Figure 2, because $E(X_i|T_i = 0)$ lies outside the distribution of covariates among those with $T_i = 1$; the observed data provide no data to fit a regression surface for $Y_{i1}$ in the vicinity of $E(X_i|T_i = 0)$, so $ACE_0$ could not be estimated in an example like this without making strong modeling assumptions that could not be tested.

If either of the regression surfaces is nonlinear—for example, as in a generalized linear model with a logarithmic or logistic link function (Cohen et al., 2003, chapter 13)— then the geometric interpretation of ACE and $ACE_1$ shown in Figure 2 no longer holds for the following reason: The average of a nonlinear function does not equal the function of the average. To obtain the ACE with nonlinear models, one would have to compute the vertical distance between the two regression surfaces for each member of the population and then average these distances over the population. The interpretation of an ACE as a vertical distance at a single point is still valid for polynomial regression models, however, because terms like $X_{i1}^2$, $X_{i1}^3$, and so forth may be regarded as additional covariates, and the regression is linear with respect to these covariates. It is also valid for regression splines, which describe the mean response using piecewise polynomial functions. The use of splines in ANCOVA is described by Little et al. (2000).

### *The Importance of Overlap*

Visualizing the ACE as in Figure 2 helps us to understand when causal inference is advisable and when it is not. The key issue is whether the distributions of the covariates among treated and untreated persons overlap sufficiently to produce stable estimates of these surfaces in regions of high covariate density (Dehejia & Wahba, 1999; Rubin, 1997). The surface $E(Y_{i0}|X_i)$ can be best estimated for values of $X_i$ near $E(X_i|T_i = 0)$; in that region, the predicted values of $Y_{i0}$ are most precise and least susceptible to bias when the relationships have been misspecified. Similarly, $E(Y_{i1}|X_i)$ can be estimated most reliably for values of $X_i$ in the neighborhood of $E(X_i|T_i = 1)$. From Figure 2, however, we see that the quality of an estimated ACE depends on how well the two surfaces can be predicted near the overall mean $E(X_i)$. If the samples are far apart in the sense that few treated individuals have covariate values resembling those of untreated individuals or vice-versa, the estimate of either surface near $E(X_i)$ may require extrapolation. Extrapolation makes an estimate of an ACE unstable and prone to bias if either regression is misspecified. Dangers of model-based extrapolation and diagnostics for assessing sensitivity to model failure are described by King (2006) and King and Zeng (2006).

Assessing the degree of overlap is easy with one or two covariates, but as $p$ increases, it becomes difficult to tell how far apart the groups really are. Comparing means for treated and untreated individuals one covariate at a time is not sufficient, because these differences do not account for correlations among the covariates. Fortunately, judging distance with many covariates becomes straightforward if we estimate propensity scores, as we describe later.

When the covariate distributions make an estimate of ACE unstable, it may still be possible to obtain a good estimate of $ACE_1$. If the study has a sufficient number of untreated persons whose covariates resemble those of treated persons—which tends to happen if the untreated sample is much larger than the treated sample—then estimates of $ACE_1$ will be based on interpolation rather than extrapolation and will be more robust. Similarly, if the covariate distribution for treated persons adequately covers the distribution for untreated persons, then estimates of $ACE_0$ will be more robust.

### *When ANCOVA Estimates an ACE*

We are now ready to describe the conditions under which the treatment effect in ANCOVA corresponds to an ACE. The treatment effect in linear ANCOVA is the coefficient $\theta$ of the treatment indicator in the regression model shown in Equation 6. The ACE is the population average difference between the potential outcomes, $E(D_i) = E(Y_{i1} - Y_{i0})$. To draw a connection between the two, we must realize that the response variable $Y_i$ in ANCOVA is either $Y_{i0}$ or $Y_{i1}$, depending on the value of $T_i$. We can write this response as $Y_i = T_iY_{i1} + (1 - T_i)Y_{i0}$, which reduces to $Y_{i1}$ if $T_i = 1$ and $Y_{i0}$ if $T_i = 0$.

If we assume that the potential outcomes are linearly related to the covariates, $E(Y_{i0}|X_i) = \alpha_0 + X_i^T\beta_0$ and $E(Y_{i1}|X_i) = \alpha_1 + X_iT\beta_1$, and—most importantly—if the treatment mechanism is unconfounded given the covariates, then

$$E(Y_i|T_i, X_i) = E(T_iY_{i1} + (1 - T_i)Y_{i0}|T_i, X_i)$$

$$= \alpha_0 + (\alpha_1 - \alpha_0)\, T_i + X_i^T \beta_0 + T_i X_i^T (\beta_1 - \beta_0). \quad (9)$$

Substituting the right-hand side of Equation 9 with $T_i = 0$ and $T_i = 1$ into Equation 7 gives

$$\theta = (\alpha_1 - \alpha_0) + X_i^T (\beta_1 - \beta_0). \quad (10)$$

Comparing Equations 10 and 8, we see that the ANCOVA parameter $\theta$ coincides with the ACE when $\beta_0 = \beta_1$, that is, when the regression planes for the potential outcomes are parallel.

If the planes are not parallel—that is, if any slope in the regression for $Y_{i0}$ does not equal the corresponding slope in the regression for $Y_{i1}$—then $\theta$ in the simple linear ANCOVA model (Equation 6) is no longer an ACE. We can make the ACE appear as a coefficient, but we will need to expand the model to include interactions. We will also need to center the covariates by subtracting from each covariate the population mean for that covariate (Aiken & West, 1991). We can do this for all covariates at once by replacing the vector $X_i$ with $(X_i - E(X_i))$. We also need to compute the product of $T_i$ with each centered covariate; the vector of these products is $T_i\,(X_i - E(X_i))$. If we expand the ANCOVA model to

$$E(Y_i | T_i, X_i) = \alpha + T_i\theta + (X_i - E(X_i))^T\beta + T_i(X_i - E(X_i))^T\eta, \quad (11)$$

where $\eta$ is a vector of interactions, then the parameter $\theta$ in this new model coincides with the ACE. If we instead center the covariates at their mean values in the *treated* population,

$$E(Y_i | T_i, X_i) = \alpha + T_i\theta + (X_i - E(X_i | T_i = 1))^T\beta + T_i(X_i - E(X_i | T_i = 1))^T\eta, \quad (12)$$

then $\theta$ coincides with $ACE_1$. (In real applications, the population means in $E(X_i)$ and $E(X_i|T_i = 1)$ will be unknown, and we will have to estimate them from the sample.)

This discussion helps us to understand the pitfalls of using the simple linear ANCOVA model (Equation 6) for causal inference. If the treatment interacts with any of the baseline measures, and those baseline-by-treatment interactions are not included in the model, then the estimated value of $\theta$ will be a biased estimate of the ACE. The potential for bias grows as we add more covariates and as the treated and untreated groups become increasingly different with respect to these covariates (Rubin, 1984, 1990). These problems may be exacerbated if one of the groups ($T_i = 1$ or $T_i = 0$) is much larger than the other. If any slopes are different in the two groups, and the appropriate interactions are not included in the model, then the estimates of these slopes will be heavily influenced by the larger group. It may be tempting to include only those Baseline $\times$ Treatment interactions that are statistically significant at some level (e.g., .05 or .10). That strategy may be dangerous, however, especially if one of the two groups is small. A small sample size in either group means that we have little power to detect Treatment $\times$ Baseline interactions. The uncertainty associated with these interactions will be large, and setting them to zero by eliminating them from the model could make the standard error for the estimate of $\theta$ unrealistically small.

Thus far, we have assumed that the ANCOVA model is linear. Nonlinear trends can be described with polynomials and regression splines (Little et al., 2000). For our purposes, those models are still "linear" because the predicted values are a linear function of a set of transformed covariates. If the responses are described by generalized linear models with nonlinear link functions, however, then the correspondence between regression coefficients and ACEs is lost. We may use generalized linear models to estimate an ACE, but it requires an extra step. We need to compute predicted values of the potential outcomes, $\hat{Y}_{i0}$ and $\hat{Y}_{i1}$, for each individual in the sample, and then average the predicted differences $\hat{Y}_{i1} - \hat{Y}_{i0}$ over the sample. The result is no longer a regression coefficient but a *regression estimate*, as we explain below.

One more point must be emphasized regarding ANCOVA. Even if the model correctly describes how the response varies with $T_i$ and $X_i$, the regression coefficient will not correspond to an ACE *if unconfoundedness does not hold*. Unconfoundedness implies that the relationships between $Y_{i1}$ and $X_i$ seen among those with $T_i = 1$, and the relationships between $Y_{i0}$ and $X_i$ seen among those with $T_i = 0$, hold for the entire population. Without this assumption, we have no basis to infer anything about missing potential outcomes. Omitting variables from ANCOVA that are related to $T_i$, even if they are not statistically significant predictors of $Y_i$, would violate unconfoundedness and may bias the causal estimate. Winship and Morgan (1999) warn against using statistical significance as the criterion for deciding whether a covariate should be included. A covariate may appear insignificant because it is correlated with $T_i$ and shares its significance with $T_i$. In other contexts, that might be a good reason to omit the covariate. For estimating a causal effect, however, that should motivate us to keep the covariate, because it is makes the unconfoundedness assumption more plausible.

### *Performance of ANCOVA in the Dieting Study*

Recall that the estimated treatment effect from linear ANCOVA in our initial sample was $-.014$ ($SE = .013$). That estimate was based on the model shown in Equation 6 with main effects for the 13 baseline variables but no interactions. We interpret it as an estimate of both ACE and $ACE_1$, because if that model were correct, the two quantities would be the same. When we centered the baseline measures at their means in the entire sample, and included all Baseline $\times$ Treatment interactions as in Equation 11, the

estimate of ACE became −.006 (*SE* = .015). When we centered the baseline measures at the sample means for the dieters, and included all Baseline × Treatment interactions as in Equation 12, the resulting estimate of $ACE_1$ was −.015 (*SE* = .013).

It is difficult to tell how well these methods actually work with only one sample, so we computed estimates, standard errors, and nominal 95% confidence intervals on the basis of a normal approximation (estimate ± 1.96 × *SE*) for all 5,000 random samples. Over repeated samples, the main-effects-only model produced an average estimated treatment effect of −.013, which lies below the true effect in the population (ACE = .003). The bias (−.013 − .003 = −.016) is small in absolute terms, but it has a large impact on confidence intervals and hypothesis tests, because in samples this large the standard errors are small. Our rule-of-thumb is that, if the magnitude of the bias in an estimate exceeds about one half of its standard error, then the performance of intervals and tests will be impaired (Collins et al., 2001). A typical standard error for this estimate is about 0.013, so its bias is larger than one standard error. Only 76.7% of the confidence intervals cover the true value. The implied Type 1 error rate (23.3%) is nearly 5 times as large as it should be. The main-effects-only model does a better job of estimating the ACE among the dieters ($ACE_1$ = −.022), but it is still noticeably biased. For that quantity, the bias is −.013 − (−.022) = .009, about 60% of a typical standard error The coverage of the nominal 95% intervals is 89.0%, and the Type 1 error rate (.11) is about twice what it should be.

ANCOVAs performance drastically improved when we included interactions. Biases in the estimates of ACE and $ACE_1$ dropped to about 30% and 40% of their standard errors, and coverage rates rose to 94.2% and 91.6%. This example underscores the need to consider Baseline × Treatment interactions when estimating causal effects from observational data. It is tempting to think that interactions are necessary only if we are investigating how treatment effects are moderated by characteristics of individuals. That is a fallacy. Failure to account for these interactions will bias an estimate of an ACE. It is also tempting to think that the interactions are necessary only if they substantially improve the fit of the model. That is another fallacy. Across our 5,000 samples, the average fit statistic for the model without interactions was $R^2$ = .341, and the average fit with interactions was $R^2$ = .344. Interactions were of little use for predicting $Y_i$, but they were crucial for estimating the ACE.

### *Additional Remarks About ANCOVA*

Our ANCOVA analyses in this simulation were not based on careful modeling. We did not check the adequacy of the models by looking for nonlinearities or heteroscedasticity. In our population, the true relationships were not linear, and the residual variance was not constant. The outcome measure was restricted to lie between 0 and 3, and the variance increased with the predicted values. If a model mistakenly assumes that the errors are homoscedastic, the estimated treated effects will not be biased. In that case, however, OLS will not be efficient, and greater precision may be obtained by switching to weighted least squares (WLS; Kutner, Nachtsheim, Neter, & Li, 2005, chapter 11). In WLS, weights should be proportional to the reciprocals of error variances. Because error variances are generally unknown, the weights will need to be derived from some combination of theory and empirical evidence. Weights should be incorporated with caution, however, because if the weights are highly variable, estimates may be unduly influenced by only a small number of individuals, and the effective sample size may be greatly reduced.

Incorrect assumptions about error variances may also distort standard errors, which could adversely affect the performance of confidence intervals and hypothesis tests. Robust or empirical standard errors that do not require correct specification of the variance are now available in many statistical packages (e.g., the GENMOD procedure in SAS). These standard errors, which are based on a sandwich formula (Liang & Zeger, 1986; White, 1980), provide valid measures of uncertainty in large samples even if the error variances are misspecified. In general, sample sizes of several hundred or more in the treated and untreated groups may be required for the robust standard errors to have the desired performance (Hardin & Hilbe, 2003).

The use of ANCOVA in observational studies has been heavily criticized by some (Berman & Greenhouse, 1992; Miller & Chapman, 2001). Results of Little et al. (2000) and our work here show that ANCOVA sometimes works well, and the method should not be categorically dismissed. Some have suggested that ANCOVA is a good choice if the distributions of baseline measures in the treated and untreated groups are similar, but if the differences between the groups are large, one should switch to a method based on propensity scores (Dehejia & Wahba, 1999; Rubin, 1997). Rubin (2001) gave a rule-of-thumb for deciding when ANCOVA adjustments are unreliable. Because that rule is based on propensity scores, we discuss them in the next section. ANCOVA versus propensity scores does not need to be an either-or decision. Many of the procedures we examine in the sections ahead will combine features of both.

Perhaps the most dangerous aspect of ANCOVA is that those who use it are often unaware of how different the groups are. Causal inference is ill-advised when the distributions of baseline variables in the groups do not sufficiently overlap, because estimated effects will then be based on extrapolation. Checking the overlap, however, is not one of the procedures usually taught in a regression course. Relationships between $T_i$ and $X_i$ are usually ignored, be-

cause ANCOVA does not model the relationships among predictors.

Variable selection was not an issue here, because our samples of $N = 6{,}000$ could accommodate all main effects and interactions without difficulty. In studies with small $N$ and many potential confounders, including all covariates and interactions may be impossible. Even if it is possible, it may be unwise, because too many superfluous predictors can make standard errors unacceptably large. When many potential confounders are present, it becomes difficult to specify a model that adequately describes all of their relationships to the outcome, because of the so-called curse of dimensionality. In those situations, we need a strategy for variable reduction. Popular variable-selection methods (e.g., stepwise regression) are not well suited to causal inference, because they are designed to reduce the mean squared error for predicting $Y_i$. For estimating an ACE, however, we need unbiased prediction of $Y_{i1} - Y_{i0}$. For that purpose, the most effective strategy for variable reduction is to estimate propensity scores.

### *Method 3: Regression Estimation*

Until now, we have assumed that a user of ANCOVA will estimate a causal effect by a regression coefficient. However, there is another way to use regression to estimate an ACE. Suppose we regress the responses for treated persons, which are values of $Y_{i1}$, on their baseline measures in the treated sample. This regression may be used to compute a predicted or fitted value $\hat{Y}_{i1}$, for any person in the study, treated or not. The average of these predictions over the entire sample estimates the population mean of $Y_{i1}$. Similarly, we may regress $Y_{i0}$ on baseline measures in the untreated sample. That model may be used to compute a prediction $\hat{Y}_{i0}$ for each person in the study, and the average of the $\hat{Y}_{i0}$s estimates the population mean of $Y_{i0}$. The difference between these averages estimates the ACE.

An average of regression predictions is called a regression estimate. It should not be confused with a regression coefficient. Under special conditions the two will agree, but they are quite different conceptually. A regression coefficient is a parameter in a model for the conditional mean of a response given a set of covariates. A regression estimate is a covariate-assisted estimate of the population mean response. Regression estimation is well-known in the survey literature (Cochran, 1977; Lohr, 1999). In a survey, $X_i$ represents frame variables that are known for all units in the population (e.g., variables used to define strata in stratified random sampling), and the response is a variable recorded only for units in the sample. In our case, the $X_i$s are covariates available for the entire sample, and the response is a potential outcome seen for only a part of the sample.

To simplify the notation, we now suppose that the vector of covariates, $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$, includes a constant term ($X_{i1} = 1$). A linear regression model for $Y_{i1}$ can be written as $E(Y_{i1}|X_i) = X_i{}^T\beta_1$, where $\beta_1$ is a vector of unknown coefficients, and the first coefficient is now an intercept. If we fit this model by applying OLS to the sample of treated persons, the vector of estimated coefficients can be written as

$$\hat{\beta}_1 = \left( \sum_i T_i X_i X_i^T \right)^{-1} \left( \sum_i T_i X_i Y_i \right). \quad (13)$$

Once again, the appearance of $T_i$ in each summand makes Equation 13 equivalent to fitting the regression model only to persons with $T_i = 1$. Similarly, a linear regression for $Y_{i0}$ can be written as $E(Y_{i0}|X_i) = X_i{}^T\beta_0$, and the vector of OLS-estimated coefficients from persons with $T_i = 0$ is

$$\hat{\beta}_0 = \left( \sum_i (1 - T_i) X_i X_i^T \right)^{-1} \left( \sum_i (1 - T_i) X_i Y_i \right). \quad (14)$$

A regression estimate for the ACE in the population may be written as

$$\hat{ACE} = \frac{1}{N} \sum_i \left( \hat{Y}_{i1} - \hat{Y}_{i0} \right), \quad (15)$$

where $\hat{Y}_{i1} = X_i^T\hat{\beta}_1$ is a prediction based on the first model, and $\hat{Y}_{i0} = X_i^T\hat{\beta}_0$ is a prediction based the second model. If the treatment mechanism is unconfounded, and if the models correctly portray the relationships between the potential outcomes and the covariates, then will consistently estimate $\hat{ACE}$. Consistency means that the bias vanishes in large samples, and that the estimate converges to the true population value as $N \to \infty$.

One may ask why Equation 15 substitutes predictions for potential outcomes that are actually seen. If we use predictions only for the missing values, the estimate becomes

$$\hat{ACE} = \frac{1}{N} \sum_i \{ T_i(Y_i - \hat{Y}_{i0}) + (1 - T_i)(\hat{Y}_{i1} - Y_i) \}, \quad (16)$$

which is an example of conditional mean imputation (Little & Rubin, 2002; Schafer & Schenker, 2000). With linear models fit by OLS, Equations 15 and 16 will be equivalent, provided that each regression model includes an intercept.

If our goal is to estimate an ACE among the treated, then a regression model for $Y_{i1}$ becomes unnecessary, because that potential outcome is observed for everyone in the treated sample. A regression estimate for $\text{ACE}_1$ is

$$\hat{ACE}_1 = \frac{\sum_i T_i (Y_i - \hat{Y}_{i0})}{\sum_i T_i}, \quad (17)$$

the average difference between $Y_{i1}$ and $\hat{Y}_{i0}$ among the treated.

Regression estimation is closely related to ANCOVA with interactions. If the same baseline measures appear as predictors in the regression models for $Y_{i1}$ and $Y_{i0}$, and if each of those models is linear, then the regression estimate of ACE (Equation 15) becomes algebraically equal to the estimate of θ in the linear ANCOVA model with Baseline × Treatment interactions (Equation 11). Under the same conditions, the regression estimate of $ACE_1$ (Equation 17) becomes equal to the estimate of θ in the ANCOVA model (Equation 12).

Despite these similarities, we have reasons to prefer regression estimates to regression coefficients for causal inference in observational studies. First, regression estimates, as we have described them, force the analyst to acknowledge that $Y_{i1}$ and $Y_{i0}$ are two different variables. Potential outcomes are not explicit in traditional applications of ANCOVA, where the models are expressed in terms of a single variable $Y_i$; the causal effects are buried by the notation, and the crucial assumptions are rarely stated. Second, by splitting the sample and fitting separate models to treated and untreated persons, all Baseline × Treatment interactions are included automatically. (If the sample size in either group is not large enough to support its own regression model, then fitting a traditional ANCOVA model with no Baseline × Treatment interactions to the pooled sample does not solve the problem but merely imposes the additional assumption of equality of slopes that cannot be evaluated.) Third, the regression estimates of Equations 15 and 17 apply very generally to all kinds of regression, including truly nonlinear models that express the means of $Y_{i1}$ and $Y_{i0}$ on a nonlinear scale. Fourth, formulas for standard errors for regression estimates, which we provide in the Appendix, do not assume that the variances of the potential outcomes have been modeled correctly. Our standard errors are robust to misspecification of mean-variance relationships, whereas the so-called model-based standard errors typically provided by linear regression software are not. In our simulation, the robust standard errors performed a little better than the model-based ones. Over 5,000 repeated samples, the coverage rates for confidence intervals from ANCOVA with interactions and model-based standard errors were 94.2% for ACE and 91.6% for $ACE_1$, but the coverage rates for intervals based on the regression estimates with robust standard errors were 94.3% for ACE and 93.2% for $ACE_1$.

## Introduction to Propensity Scores

### *Definition and Estimation*

Rosenbaum and Rubin (1983) defined the propensity score as $P(T_i = 1|X_i, Y_{i0}, Y_{i1})$, the conditional probability of receiving the treatment given the covariates and potential outcomes. When the treatment mechanism is unconfounded, the propensity score is unrelated to $Y_{i0}$ and $Y_{i1}$ after $X_i$ is taken into account. So we may drop the potential outcomes from this notation and express the propensity score as a function of $X_i$ alone. We write the propensity score for one individual as $\pi(X_i)$ or, more simply, as $\pi_i$.

In an observational study, propensity scores are unknown and must be estimated from the sample. The most common way to accomplish this is by logistic regression (Hosmer & Lemeshow, 2003; Menard, 2002). Let $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ denote the vector of covariates for individual *i*, which we suppose includes a constant ($X_{i1} = 1$). The logistic model assumes $\log [\pi_i/(1 - \pi_i)] = X_i^T \gamma$, where $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p)^T$ is a vector of regression coefficients that are usually estimated by maximum likelihood. Solving for $\pi_i$ gives

$$\pi_i = \frac{exp(X_i^T\gamma)}{1 + exp(X_i^T\gamma)}, \tag{18}$$

and estimated propensity scores are obtained by plugging the estimated γs into this formula.

The popularity of the logistic model stems from the fact that estimation software is widely available. In addition, when the coefficients are of interest, the slopes $\gamma_2, \ldots, \gamma_p$ may be exponentiated and interpreted as odds ratios. However, our main purpose in fitting a propensity model is not to interpret coefficients but to obtain predicted propensities $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_N$. Predictions from a logistic model may be unduly influenced by outliers (cases having $\hat{\pi}_i \approx 0$ and $T_i = 1$, or $\hat{\pi}_i \approx 1$ and $T_i = 0$; Pregibon, 1982). Albert and Chib (1993) and Liu (2004) have advocated a more robust version called the robit model, which replaces the logistic function in Equation 18 by the cumulative distribution function for the Student *t*-distribution with $\nu$ degrees of freedom. Kang and Schafer (2007) have reported improved performance in one example by estimating propensity scores under a robit model with $\nu = 4$, but software for robit modeling is not widely available. Some have proposed to estimate propensity scores without strong parametric assumptions using classification trees (Breiman, Friedman, Olshen, & Stone, 1984; Luellen, Shadish, & Clark, 2005; McCaffrey, Ridgeway, & Morral, 2004) and neural networks (King & Zeng, 2002). In our simulated study, we estimate propensity scores by logistic regression. This approach gives adequate performance in this context, but alternatives to the simple logistic model are certainly worth considering.

### *Why Use Propensity Scores?*

Propensity scores are used in many ways. Yet a question that deserves to be asked is why do we need them at all? Similar questions have been raised over whether to use

sampling weights in the analysis of survey data (Chambers & Skinner, 2003; DuMouchel & Duncan, 1983; Pfefferman, 1993, 1996). The survey weight for a sampled individual is one divided by the probability that the individual was selected into the sample, and is analogous to $1/\pi_i$. In the missing-data literature, debates occur over whether one should model probabilities of missingness or simply ignore them (Little & Rubin, 2002). Propensity scores play the same role as probabilities of missingness, because the $T_i$ is simply an indicator that tells us which of the two potential outcomes is missing.

The answer to this question depends on what assumptions we are willing to make about the distributions of the potential outcomes. For analyses based on parametric models for $Y_{i1}$ and $Y_{i0}$, the propensity scores become irrelevant as long as the models are correct and the treatment assignment is unconfounded. If we believe that we can accurately model the potential outcomes and their relationships to the covariates, then there is no need to use propensity scores to estimate an ACE. Methodologies for causal inference spawned by Rosenbaum and Rubin (1983), on the other hand, model the $T_i$s but make no assumptions about $Y_{i0}$ or $Y_{i1}$. Those methods are in part a reaction to misapplications of ANCOVA, where analysts were often unaware of the sensitivity of their results to model failure. Work by Rubin (1973, 1979, 1997) suggests that ANCOVA is effective when the distributions of covariates among the treated and untreated are similar, but propensity-score methods are better when the groups are very different. In a few examples, propensity scores have been shown to do a better job than ANCOVA at reproducing experimental effects (Dehejia & Wahba, 1999; Lalonde, 1986). On the other hand, one should not believe that propensity-score methods are inherently more robust than models for the outcomes, because overreliance upon a misspecified propensity model may produce large biases as well (Kang & Schafer, 2007). Shadish, Clark, and Steiner (in press) have presented a unique example that combines a randomized experiment with an observational study, and they have argued that the choice between outcome model-based methods and propensity model-based methods is less important than the subjective choice of which covariates to include in the adjustment.

Recognizing that modeling the propensities and modeling the potential outcomes can both be advantageous, some have sought a middle ground that combines the two approaches in a single procedure. Elliot and Little (2000) and Little and An (2004) have applied regression models that allow the response to vary as a function of the propensity scores, which helps to mitigate some of the dangers of model failure. Dual modeling of responses and propensities is also an underlying theme of the semiparametric estimation theory of Robins et al. (1994, 1995), Rotnitzky et al. (1998), and van der Laan and Robins (2003). Using two models, they have constructed estimates that are doubly robust in the sense that they consistently estimate the ACE if either model is true. Doubly robust estimators are discussed later in this article. They are not a panacea, however, and they may not outperform single-model strategies when both models are wrong (Kang & Schafer, 2007).

### *Estimated Propensity Scores in the Dieting Study*

Using our initial sample of $N = 6{,}000$, we fit a logistic model to predict girls' propensities to diet from the 13 baseline variables (main effects only). Most of the effects were statistically significant, which is not surprising given the large sample size. The most significant predictor was self-rating of weight (SLFWGHT). Girls who consider themselves overweight were more likely to diet. The next most significant predictor was self-liking (LIKESLF). Girls who reported that they disliked themselves were more likely to diet. The directions of the effects were consistent with the results of previous studies of adolescent dieting behavior (Stice, Mazotti, Krebs, & Martin, 1998). We do not report coefficients, standard errors, or significance levels, however, because the significance of individual predictors is a poor criterion for judging the usefulness of a propensity model. The purpose of the model is to predict the $\pi_i$s and use the predictions to reduce selection bias arising from nonrandom treatment assignment.

One should not assume that the logistic link is correct merely because it is convenient. To test the link function, Hinkley (1985) suggested computing the linear predictor $X_i^T\hat{\gamma}$ for each individual, creating a new covariate equal to the squared linear predictor, and refitting the model including this new covariate. A significant coefficient for the new covariate suggests that the link function is inadequate. In our sample, the additional covariate is not significant ($p = .93$), so a careful analyst would have little reason to question the logistic form.

Our logistic model, however, is not correct. Propensities to diet in the population were created by a nonlogistic binary regression with many interactions. In particular, the logistic model does not accurately represent the true propensities for those who are least likely to diet. The smallest fitted propensity in the sample is $\hat{\pi}_i = .004$, but the smallest true propensity is $\pi_i = .010$. We will see the effects of inaccuracy in predicting values of $\pi_i$ near zero, a problem that negatively impacts certain classes of propensity-based procedures.

### *Propensity Scores and Balance*

The key property of propensity scores that makes them useful for causal inference is that they balance the distributions of the covariates (Rosenbaum & Rubin, 1983). In theory, treated and untreated persons with identical propensity scores should have identical distributions for the co-

variates that were used to estimate the propensities. In mathematical terms, this balancing property is

$$P(X_i | \pi_i = c, T_i = 1) = P(X_i | \pi_i = c, T_i = 0) \quad (19)$$

for all values of $c$ between 0 and 1. In a randomized experiment, the distributions at baseline are only randomly different, which allows us to make an unbiased comparison. For an observational study, Equation 19 indicates that if we divide the population into groups of constant propensity, then the study will resemble a randomized experiment within each group. The average difference in response between treated and untreated persons within a group of constant propensity is an unbiased estimate of the ACE for that group.

Because we assume that the treatment is unconfounded given $X_i$, the propensity is a function of the measures in $X_i$. *The propensity score is a baseline variable, and it can be used as any other baseline variable*. It is an extremely useful variable, because it captures in one dimension the essence of how the groups differed when the study began. The propensity score (or any transformation of it, such as the logit propensity score) is the simplest summary of $X_i$ that can achieve balance with respect to all the covariates (Rosenbaum & Rubin, 1983).

The ability to balance covariate distributions is a property of the *true* propensity scores. In practice, we must use propensities estimated from a model. If the model is accurate, it helps us to form groups that are comparable enough to be handled as if they had come from a randomized experiment. To illustrate, we estimated propensities for the 6,000 girls in our sample using the logistic model. We then examined the 329 girls whose estimated propensities fell in the interval [0.35, 0.40], which was chosen arbitrarily. We then performed a *t*-test on each baseline measure to compare the means for dieters and nondieters, first in the overall sample and then in the subgroup. Results from these tests are shown in Table 4. In this table, $\bar{x}_0$ and $\bar{x}_1$ denote the means for nondieters and dieters, respectively, and $T$ is the statistic from the pooled *t*-test. Of the 13 comparisons, 11 are significant at the .05 level in the full sample, but none are significant in the subgroup. Power to detect differences is much greater in a sample of 6,000 than in a sample of 329. To eliminate dependence on sample size, we computed Cohen's effect size measure $d$, defined as $(\bar{x}_1 - \bar{x}_0)/S$. where $S$ is the pooled standard deviation. Large values of $d$ in the full sample become much smaller in the subgroup. The average value of $|d|$ across all 13 comparisons in the full sample is 0.23, but in the subsample it is only 0.07. Within this subgroup, we no longer see evidence of systematic differences between dieters and nondieters when we examine the baseline measures one at a time. Comparisons of one variable at a time are not sufficient to establish balance, however, because they do not take into account the correlations among the covariates (Cochran, 1965). One should also compare the distributions of the estimated propensity scores, as we now describe.

### *Comparing the Distributions of Propensity Scores*

Propensity scores clarify how different the treated and untreated groups were at the beginning of the study. The estimated propensity score, or any monotonic transformation of it, is the baseline variable that maximally discriminates between the groups. It is more generally applicable than the linear discriminant function, because it does not assume that the covariates share a common within-group covariance structure. If the $\pi_i$s are modeled by logistic regression, then it is natural to examine the logit-propensity, $log[\hat{\pi}_1/(1 - \hat{\pi}_i)]$, which is the estimated log-odds of treatment. From Equation 18, we see that this value can also be expressed as $X_i^T\hat{\gamma} = \hat{\gamma}_1 X_{i1} + \ldots + \hat{\gamma}_p X_{ip}$, where the $\hat{\gamma}$s are the estimated coefficients. In texts on logistic regression, the estimated logit-propensity is known as the linear predictor. Stacked histograms or side-by-side box plots of the estimated propensity scores or linear predictors for the two groups will quickly reveal the degree of overlap.

Histograms of the estimated propensities and linear predictors from our initial sample are shown in Figure 3. The distributions of the propensities are clearly different, but their ranges look similar. Nondieters have $M = 0.17$, $SD = 0.13$, minimum = 0.004, and maximum = 0.83; dieters have $M = 0.32$, $SD = 0.17$, minimum = 0.009, and maximum = 0.80. On the logit scale, however, the ranges look less similar, because differences in the tail where $\hat{\pi}_i \approx 0$ are magnified. The logit propensities for nondieters have $M = -1.86$, $SD = 1.01$, minimum = $-5.45$, and maximum = 1.58, whereas the dieters have $M = -0.89$, $SD = 0.96$, minimum = $-4.74$, and maximum = 1.38. It is usually more appropriate to compare the distributions for the linear predictors, because on the logit scale the distributions tend to be less skewed, and the variances tend to be more nearly equal (Rubin, 2001).

These plots help us to assess overlap. Overlap refers to how well the distribution for each group covers that of the other group. The logit-propensities for nondieters, who comprise 80% of the sample, effectively span those of the dieters. This bodes well for estimating $ACE_1$. For any girl who dieted, we can easily find another girl who was about as likely to diet but did not, and who can serve as a proxy to help us infer what the dieter's missing value of $Y_{i0}$ might be. If the propensity model is correct, then dieters and nondieters with the same propensities may be regarded as participants in a randomized experiment whose treatments were not self-selected but decided by the flip of a coin.

The logit-propensities for the dieters, however, do not cover those of the nondieters very well at the low end. Of the nondieters, 20% had logit propensities below $-2.58$

Table 4
*Mean Comparisons for Baseline Measures Among Nondieters and Dieters From N = 6,000 Girls in the Full Simulated Sample, and Among the 329 Girls With Estimated Propensity Scores $\hat{\pi}_i$ Between 0.35 and 0.40: Mean for Nondieters ($X_0$), Mean for Dieters ($X_1$), Cohen's Effect Size d, and T Statistic*

| | | Full sample | | | | Group with $.35 \leq \hat{\pi}_i \leq .40$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| Name | Description | $x_0$ | $x_1$ | $d$ | $T$ | $x_0$ | $x_1$ | $d$ | $T$ |
| DISTR.1 | Emotional distress at Wave I (minimum = 0, maximum = 2.84) | 0.62 | 0.71 | 0.22 | 6.82 | 0.81 | 0.81 | 0.01 | 0.09 |
| BLACK | 1 = Black, 0 = otherwise | 0.26 | 0.17 | −0.19 | −5.90 | 0.09 | 0.08 | −0.03 | −0.23 |
| NBHISP | 1 = non-Black Hispanic, 0 = otherwise | 0.15 | 0.15 | 0.02 | 0.65 | 0.16 | 0.19 | 0.07 | 0.62 |
| GRADE | Grade in school at Wave I (7, . . . , 11) | 9.16 | 9.37 | 0.15 | 4.69 | 9.60 | 9.60 | 0.01 | 0.05 |
| SLFHLTH | Self-rating of overall health (1 = *excellent*, 2 = *very good*, 3 = *good*, 4 = *fair*, 5 = *poor*) | 2.20 | 2.35 | 0.17 | 5.29 | 2.57 | 2.52 | −0.05 | −0.41 |
| SLFWGHT | Self-rating of weight (1 = *very underweight*, 2 = *slightly under*, 3 = *about right*, 4 = *slightly over*, 5 = *very over*) | 3.19 | 3.84 | 0.88 | 27.29 | 4.01 | 4.01 | 0.03 | 0.24 |
| WORKHARD | "When you get what you want, it's usually because you worked hard for it" (1 = *strongly agree*, . . . , 5 = *strongly disagree*) | 2.14 | 2.05 | −0.09 | −2.85 | 1.97 | 1.96 | −0.02 | −0.19 |
| GOODQUAL | "You have lots of good qualities" (1 = *strongly agree*, . . . , 5 = *strongly disagree*) | 1.80 | 1.84 | 0.06 | 1.75 | 1.87 | 1.99 | 0.18 | 1.58 |
| PHYSFIT | "You are physically fit" (1 = *strongly agree*, . . . , 5 = *strongly disagree*) | 2.24 | 2.53 | 0.32 | 9.99 | 2.73 | 2.77 | 0.05 | 0.43 |
| PROUD | "You have a lot to be proud of" (1 = *strongly agree*, . . . , 5 = *strongly disagree*) | 1.76 | 1.86 | 0.13 | 3.96 | 1.95 | 2.09 | 0.16 | 1.44 |
| LIKESLF | "You like yourself just the way you are" (1 = *strongly agree*, . . . , 5 = *strongly disagree*) | 2.09 | 2.52 | 0.43 | 13.38 | 2.80 | 2.88 | 0.08 | 0.70 |
| ACCEPTED | "You feel socially accepted" (1 = *strongly agree*, . . . , 5 = *strongly disagree*) | 2.14 | 2.35 | 0.21 | 6.56 | 2.43 | 2.55 | 0.12 | 1.08 |
| FEELLOVD | "You feel loved and wanted" (1 = *strongly agree*, . . . , 5 = *strongly disagree*) | 1.78 | 1.93 | 0.18 | 5.48 | 2.05 | 2.17 | 0.13 | 1.18 |

(corresponding to $\hat{\pi}_i < 0.070$), but only 3.9% of the dieters (48 girls in the sample) did. In this region, there are few observed values of $Y_{i1}$, so inferences about the effects of dieting among girls who are unlikely to diet will rest heavily on modeling assumptions. In other applications, one may find that in parts of the propensity distribution for one group, there may be no corresponding individuals in the other group. In that case, there would be little scientific basis for estimating an ACE for the whole population, because in the region without overlap the inference would be based entirely on extrapolation. Comparisons should then be limited to subgroups whose propensities lie in the region of overlap.

Logit-propensities also help us to quantify imbalance. In this example, the means of the logit-propensities for nondi-

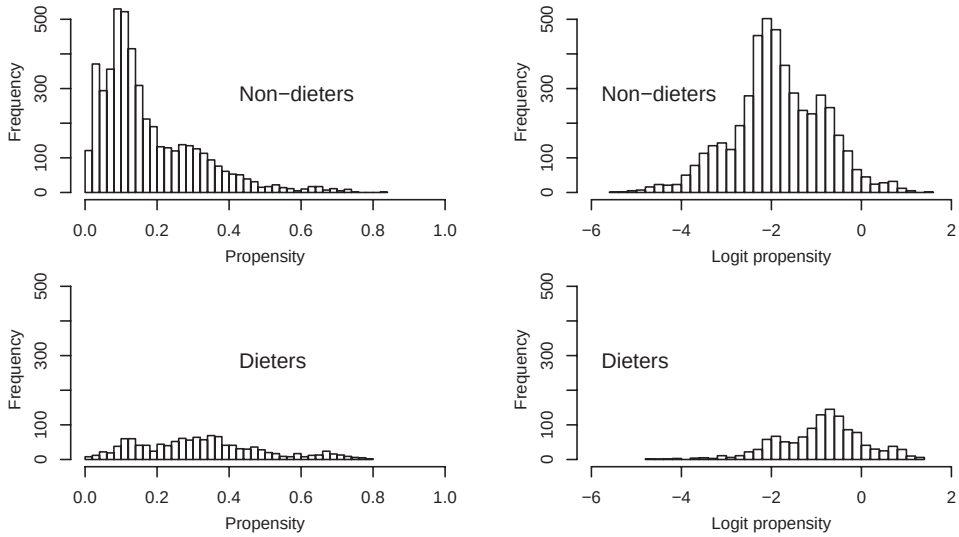


*Figure 3.* Histograms of estimated propensity scores and linear predictors (logit-propensities) in a sample of $N = 6{,}000$ girls from the synthetic population. The two panels on the left-hand side compare the distributions of estimated propensity scores for dieters and nondieters. The two panels on the right-hand side compare the distributions of linear predictors for dieters and nondieters.



eters and dieters differ by about one standard deviation. Rubin (2001) has given a rule-of-thumb that if the difference between the means of the logit-propensities exceeds half of the pooled within-group standard deviation, then causal inferences based on ANCOVA are not trustworthy. This rule—which was apparently intended for models without baseline-by-treatment interactions—helps to explain why ANCOVA without interactions did not perform well over repeated samples for estimating ACE in this example.

### *Guidelines for Constructing a Propensity Model*

In typical applications of logistic regression, analysts remove predictors whose coefficients are not significantly different from zero. For propensity modeling, however this practice is strongly discouraged. Prediction and fit, not parsimony, are the criteria by which the model should be judged. Fit statistics for logistic regression, including analogues of multiple linear regression $R^2$, are reviewed by DeMaris (2002). Overfitting a propensity model can be beneficial, because propensities estimated by a rich model may carry more information about the potential outcomes in the sample than the true propensity scores, producing more efficient estimates of ACEs (Lunceford & Davidian, 2004; Rubin & Thomas, 1996). To increase the efficiency of a propensity-based method, it makes sense to include covariates in the propensity model that are related to the outcome variable $Y_i$, whether they are significantly related to $T_i$ (Kang & Schafer, 2007; Vartivarian & Little, 2002).

When many potential confounders are available, it may be necessary to sift through them to identify a subset to include in the propensity model. In other types of modeling, selecting predictors from a large pool increases the chance of Type 1 errors unless adjustments (e.g., Bonferroni corrections) are made. This is not a problem for propensity modeling, however, because potential outcomes do not appear in the model. Propensity scores should be estimated before outcomes are examined, so causal conclusions need not be weakened by suspicions that they arose by chance from dredging or mining the data (Rubin, 2001).

For selecting baseline variables to include in a propensity model, we offer the following advice. Include any variable that, for scientific reasons, is thought to be an important predictor of $T_i$. Include any variable that is thought to be an important predictor of $Y_i$. Finally, include any additional variable that is even mildly significant ($p > .10$ or .15) as a predictor of $T_i$ in the sense of a bivariate association, apart from any other predictor. This advice is most appropriate for large samples, where models with many covariates are not problematic. In smaller samples, including too many covariates will eventually separate the distributions of estimated propensities to the extent that they will no longer overlap. In those situations, selection of covariates will need to be carefully guided by theory and subject-matter knowledge about how participants may have come to receive the treatments that they had.

## Using Propensity Scores to Estimate ACEs

### *Method 4: Matching*

Propensity scores are used to select matched subsamples of treated and untreated persons whose covariate distribu-

tions are similar enough that selection bias is no longer a major issue (Rosenbaum, 2002). Matching works best when one of the two groups—typically the control—is substantially larger than the other. For each member of the smaller group, someone in the larger group with similar pretreatment characteristics is selected. The excess members of the larger group are discarded, and the analysis proceeds as if the two groups are comparable. At first glance, this strategy may seem wasteful, because data from a large number of participants may be ignored. In a two-sample comparison of means, however, precision is largely driven by the sample size of the *smaller* group, so discarding the excess individuals typically produces only a slight reduction in power (Cohen, 1988).

In this discussion, we focus on 1:1 matching, in which each person in the smaller group is matched to one person in the larger group. When the groups are very different in size, it may be worthwhile to consider procedures in which a variable number of similar persons from the larger group are found for each person in the smaller group (Ming & Rosenbaum, 2000).

The key issue in matching is how to identify persons who are similar when $X_i$ is high-dimensional. The balancing property (Equation 19) suggests that similarity should be judged by the propensity score. Matching on propensity scores mimics the results of a randomized block experiment in which individuals within blocks of constant propensity are randomly assigned to $T_i = 0$ or $T_i = 1$. If individuals are also matched on additional components or summaries of $X_i$, those summaries play the role of additional blocking factors.

A carefully executed matching procedure can eliminate most of the bias due to nonrandom treatment assignment related to the set of measured covariates in $X_i$. After the matching is completed, the matched samples may be compared by an *unpaired t*-test. ("Matching" erroneously suggests that the resulting data should be analyzed as if they were matched pairs. The treated and untreated samples should be regarded as independent, however, because there is no reason to believe that the outcomes of matched individuals are correlated in any way.) Even after matching, small random differences in the distributions of covariates between the treated and untreated groups may remain. We can adjust for this remaining imbalance by ANCOVA (Rubin & Thomas, 2000). ANCOVA applied to propensity-matched subsamples serves the same purpose as it does in a randomized experiment: to reduce variability and increase the power of the comparison. ANCOVA may also help to remove residual bias when the matching is less than perfect.

Techniques for propensity-score matching are reviewed by D'Agostino (1998), Rosenbaum (2002), and Rubin and Thomas (1996). One popular procedure is Mahalanobis-metric matching within calipers defined by the logit-propensity score. Let $\hat{\eta}_i = log[\hat{\pi}_i/(1 - \hat{\pi}_i)]$ denote the estimated logit-propensity. For each individual in the smaller group, a pool of potential matches in the larger group is identified whose logit-propensities are in the interval $\hat{\eta}_i \pm c$, where $c$ is typically one quarter of the within-group standard deviation of $\hat{\eta}_i$. Within this pool, the closest individual in terms of Mahalanobis distance is selected. This algorithm has been implemented in a SAS macro called GREEDY (Parsons, 2000), a Stata module called PSMATCH2 (Leuven & Sianesi, 2003), and an R package called MatchIt (Ho, Imai, King, & Stuart, 2004).

Matching sometimes makes it difficult to estimate an ACE for the desired population. If the smaller group consists of treated persons, and every member of this group is matched to an untreated person, then an estimate from the matched samples will represent $ACE_1$. If no matches can be found for some treated persons, then the estimated causal effect will only describe the subpopulation of treated persons whose propensities overlap with the untreated.

In our simulated study, we applied Mahalanobis-metric matching on $X_i$ within logit-propensity calipers as recommended by Rosenbaum and Rubin (1985). Of the 1,220 dieters in our initial sample, 1,194 were successfully matched to nondieters, and the remaining 26 were discarded. An unpaired *t*-test applied to the matched samples yielded an estimated mean difference of .009 (pooled $SE = .019$), which we interpret as an estimate of $ACE_1$. Matching followed by a *t*-test did not perform well over the 5,000 repeated samples. The bias in the estimates (.035) is about twice the size of a typical standard error, and only 58.6% of the confidence intervals covered the true value of $ACE_1$.

Why did this procedure perform so poorly? It is because the matched samples are still not perfectly balanced with respect to the covariates. We refit our logistic model to the matched sample and compared the estimated logit-propensities for the two groups. The means were 15% of a standard deviation apart, which is statistically significant ($p < .01$). A discrepancy of this size is still large enough to wreak havoc on the performance of a *t*-test in a large sample. It is, however, well below Rubin's (2001) cutoff to be reliably corrected by simple linear ANCOVA. We applied linear ANCOVA with 13 covariates and no interactions to the matched samples, and the bias in the resulting estimate became unimportant (about one quarter of a typical standard error). Of the intervals, 94.3% covered the true $ACE_1$.

Matching helps to reduce bias. However, as this example shows, matching may not be enough, and covariance adjustments on matched samples are still recommended. The propensities in our matching procedure were based on a logistic model that was incorrect. As we have noted, the model failed to predict propensities accurately in the region where $\pi_i \approx 0$. Failure in this region did little damage to the performance of matching, however, because this region consists largely of unmatched nondieters who were trimmed away.

### *Method 5: Inverse-Propensity Weighting*

Weights are used in sample surveys to adjust for unequal probabilities of selection (Horvitz & Thompson, 1952; Lohr, 1999). If some types of individuals are sampled at higher rates than others, observations from the oversampled groups must be deemphasized, and observations from the undersampled groups must be accentuated, to obtain an unbiased estimate for the population. Consider a national survey in which persons of Hispanic origin are sampled at a rate of 1:20,000, and others are sampled at a rate of 1:50,000. Let $Y_i$ be a measurement for participant $i$, and let $w_i = 20{,}000$ if the person is Hispanic and $w_i = 50{,}000$ otherwise. The weighted average $\Sigma_i w_i Y_i / \Sigma_i w_i$ would be an unbiased estimate of the population mean of $Y_i$. The weight $w_i$ is the number of population persons represented by person $i$. It is also the reciprocal of the probability with which he or she was sampled.

The same principle can be used with potential outcomes to estimate a population ACE. The values of $Y_{i1}$ that we actually see are a nonrepresentative selection of those in the full sample, chosen with probabilities equal to the propensity scores $\pi_1, \ldots, \pi_N$. An unbiased estimate of $E(Y_{i1})$ is $\Sigma_i w_i Y_i / \Sigma_i w_i$, where $w_i = 1/\pi_i$, and the sums are taken over individuals with $T_i = 1$. Similarly, the values of $Y_{i0}$ that we see are a nonrepresentative subsample chosen with probabilities $(1 - \pi_i)$. An unbiased estimate of $E(Y_{i0})$ is $\Sigma_i w_i^* Y_i / \Sigma_i w_i^*$, where $w_i^* = 1/(1 - \pi_i)$, and the sums are over individuals with $T_i = 0$. The difference between these weighted averages is an unbiased estimate of ACE $= E(Y_{i1}) - E(Y_{i0})$.

To compute this estimate, we need to replace the unknown $\pi_i$s by estimates from a propensity model. The inverse-propensity weighted (IPW) estimate may be written as

$$\hat{ACE} = \frac{\sum_i T_i \hat{\pi}_i^{-1} Y_i}{\sum_i T_i \hat{\pi}_i^{-1}} - \frac{\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1} Y_i}{\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1}}, \quad (20)$$

where $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$ is the observed outcome. The sums in Equation 20 are taken over the full sample, because irrelevant terms are wiped out by the factors $T_i$ and $(1 - T_i)$.

Similar principles may be used to estimate $\text{ACE}_1 = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1)$. The term $E(Y_{i1}|T_i = 1)$ may be estimated by $\Sigma_i T_i Y_i / \Sigma_i T_i$, the unweighted mean response among the treated. To estimate $E(Y_{i0}|T_i = 1)$, however, we must weight the untreated persons to represent the *treated* population. For that purpose, the weight assigned to $Y_{i0}$ should be $\pi_i/(1 - \pi_i)$ (Hirano & Imbens, 2001). The IPW estimate of $\text{ACE}_1$ is

$$\hat{ACE}_1 = \frac{\sum_i T_i Y_i}{\sum_i T_i} - \frac{\sum_i (1 - T_i) \hat{\pi}_i (1 - \hat{\pi}_i)^{-1} Y_i}{\sum_i (1 - T_i) \hat{\pi}_i (1 - \hat{\pi}_i)^{-1}}. \quad (21)$$

IPW makes no assumptions about how the potential outcomes are distributed or how they are related to covariates. Consistency of IPW estimates does require a correctly specified model for the propensity scores.

Standard errors for IPW estimates should reflect the fact that the $\pi_i$s have been replaced by $\hat{\pi}_i$s. Methods for computing standard errors that account for estimation of the $\pi_i$s were developed by Robins et al. (1995). Surprisingly, using estimated rather than true propensities tends to *reduce* variability. The $\hat{\pi}_i$s often convey more information than the true $\pi_i$s, especially if the propensity model has been overfitted (Lunceford & Davidian, 2004; Rubin & Thomas, 1996). Formulas for standard errors that assume the propensities are estimated by logistic regression are provided in the Appendix.

Applying IPW to our first sample, we obtain $\hat{ACE} = -.005$ ($SE = .028$). Over repeated samples, this procedure is biased and inefficient. The root-mean-square error (RMSE), computed from the 5,000 sample estimates $\hat{ACE}^{(j)}$ by

$$RMSE = \sqrt{\frac{1}{5{,}000} \sum_{j=1}^{5{,}000} (\hat{ACE}^{(j)} - ACE)^2},$$

is a measure of imprecision that combines variance and bias. The RMSE of the IPW estimate is nearly 70% larger than that of ANCOVA with interactions. We can see why by examining the weights. The weights $1/(1 - \hat{\pi}_i)$ for the nondieters range from 1.00 to 5.87, but the weights $1/\hat{\pi}_i$ for the dieters range from 1.25 to 115.4. The 61 dieters with the largest weights—who compose only 5% of the dieters in the sample—account for nearly 28% of their total weight. IPW estimates are greatly influenced by observations with $T_i = 1$ but $\hat{\pi}_i \approx 0$, and by observations with $T_i = 0$ but $\hat{\pi}_i \approx 1$. These outliers are valuable for estimating causal effects. Dieters with low propensity to diet provide excellent proxy information for predicting the missing $Y_{i0}$s, and nondieters with high propensity are good proxies for the missing $Y_{i1}$s. The problem with IPW is that it relies on these outliers too much. Placing undue weight on a few individuals leads to estimates with high variance. IPW is also susceptible to bias from an incorrectly specified propensity model; small errors of prediction in the tails where $\pi_i \approx 0$ and $\pi_i \approx 1$ may lead to very large errors in the weights (Kang & Schafer, 2007).

IPW fared better for estimating $\text{ACE}_1$, because the weights $\hat{\pi}_i/(1 - \hat{\pi}_i)$ applied to the nondieters in that estimate were not extreme. Even for $\text{ACE}_1$, however, IPW was less efficient than other methods described in this article. IPW can be inefficient because it is based only on propensity scores. Propensities help to reduce selection bias, but they may not effectively convey what is really known about the missing potential outcomes. Suppose, for example, that a baseline measure in $X_i$ is highly correlated with the out-

come but is unrelated to $T_i$. This covariate would contribute little to the $\hat{\pi}_i$s, because its coefficient in the propensity model would be nearly zero. It would, however, be valuable for predicting the missing values of $Y_{i0}$ and $Y_{i1}$. Without a model for the potential outcomes, predictive information in covariates related to the outcomes can be difficult to recover.

### *Method 6: Subclassification*

Propensity scores are also used to define subclasses or strata in which treated and untreated individuals may be compared without bias (Rosenbaum & Rubin, 1983). The balancing property (Equation 19) implies that if we are able to create classes in which the propensity scores are homogeneous, the treated and untreated persons within each class may be compared as if they had participated in a completely randomized experiment. Suppose we assign participants to a set of ordered classes $s = 1, \ldots, S$ based on their values of $\hat{\pi}_i$. Let $\hat{\theta}_s$ denote an estimate of the ACE in class $s$, and let $\hat{V}_s = \hat{V}ar(\hat{\theta}_s)$ denote its estimated variance. Each $\hat{\theta}_s$ may be a simple difference in means, but we may also apply ANCOVA within classes if the sample sizes are large enough. The subclassified estimate of the overall ACE is $\hat{ACE} = \sum_s (N_s/N)\hat{\theta}_s$. where $N_s$ is the sample size in class $s$, and $N$ is the total sample size. This is the weighted average of the class-specific treatment effects, weighted by the proportions of individuals falling into each class. The estimated variance of $\hat{ACE}$ is $\sum_s (N_s/N)^2 \hat{V}_s$. In a similar way, we can estimate the ACE among the treated if we weight each $\hat{\theta}_s$ by the proportion of *treated* persons falling into that class. Let $N_s$* be the number of treated persons in class $s$, and let $N$* be the total number of treated persons. The subclassified estimate of $ACE_1$ is $\sum_s (N_s^*/N^*)\hat{\theta}_s$, and its estimated variance is $\sum_s (N_s^*/N^*)^2 \hat{V}_s$.

Subclassified estimates may be viewed as IPW estimates in which the weights have been coarsened, that is, grouped together into a few categories and approximated in each category by a constant. Coarsening stabilizes the weights and mitigates the impact of model failure when some propensities are near zero or one (Kang & Schafer, 2007; Little & Rubin, 2002).

To form the classes, Rosenbaum and Rubin (1984) have suggested using five groups of approximately equal size, with boundaries at the 20th, 40th, 60$^{th}$, and 80th percentiles of the estimated propensity scores. This rule is based on an argument by Cochran (1968) that five classes remove about 90% of the selection bias from an estimate of a population mean. Subclassified estimation requires at least a few treated and untreated persons in each class. If the five-group method does not yield enough persons to compute $\hat{\theta}_s$ and $\hat{V}_s$ in all classes, it shows that the propensity distributions for treated and untreated persons do not overlap sufficiently to reliably estimate the ACE for the full population. Depending on how the $\hat{\pi}_i$s are distributed, some classes may cover a wide range of propensities, which may cause the estimates within those classes to be biased. Bias can be reduced by splitting these classes, provided that a sufficient number of treated and untreated persons remains in each class.

After the classes are formed, we must decide how to estimate the ACEs within them. The simplest way is to compute a simple difference in means, with a standard error (pooled or unpooled) obtained in the usual way. We may also fit an ANCOVA model within each class, using some or all of the covariates in $X_i$. If we could form classes perfectly so that the $\pi_i$s in each class were constant, the distributions of covariates for the treated and untreated in each class would still be randomly different; to the extent that these covariates are correlated with the response, adjusting for the random differences is beneficial.

Using our first sample, we ordered the participants by $\hat{\pi}_i$. divided them into five classes of 1,200, and estimated the ACE in each class by a difference in means. We also estimated the ACEs by linear ANCOVA (Equation 6) adjusting for one covariate, the baseline value of emotional distress. Estimates and pooled standard errors within classes are shown in Table 5. The ANCOVA estimates have smaller standard errors than the mean differences, suggesting that the covariance adjustment increased precision. The class-specific estimates show an interesting trend. In Classes 1–3, where propensities to diet are below average, the estimated effects are positive. Although these effects are not statistically significant, they suggest that dieting tends to increase distress for girls in these classes. In Classes 4 and 5, where propensities to diet are higher, the estimated effects are negative, suggesting that dieting tends to lower distress for girls in these classes. Trends like this, which are not uncommon in observational studies, indicate that those who are more likely to accept the treatment are those for whom it is more beneficial. If individuals could predict whether the treatment is beneficial for them, and if they were behaving rationally, then we would expect this pattern to emerge (Manski, 1999).

The results in Table 5 suggest that our class definitions can be improved. Classes 4 and 5 cover wide ranges of propensities and contain large numbers of dieters and nondieters, so we can afford to split them into groups that more homogeneous. We divided Class 4 into two groups of 600 and Class 5 into four groups of 300. Across the now nine classes, the minimum number of nondieters is 136, and the minimum number of dieters is 61. Combining the mean differences across the nine classes gives $\hat{ACE} = .005$ ($SE = .019$), and $\hat{ACE}_1 = -.024$ ($SE = .017$). Combining the ANCOVA estimates gives $\hat{ACE} = -.004$ ($SE = .015$), and $\hat{ACE}_1 = -.023$ ($SE = .015$). Over repeated samples, this method performs better than IPW; it has low bias, high efficiency, and accurate coverage.

Table 5
*Estimated Average Causal Effects of Dieting on Emotional Distress in the Full Simulated Sample of N = 6,000 Adolscent Girls, Estimated Within Propensity Classes by (a) an Unadjusted Mean Difference and (b) ANCOVA Adjusting for Emotional Distress at Baseline*

| | Range ($\hat{\pi}_i$) | | $N$ | | | (a) Difference | | (b) ANCOVA | |
|---|---|---|---|---|---|---|---|---|---|
| Class | minimum | maximum | $T_i = 0$ | $T_i = 1$ | Total | Estimate | *SE* | Estimate | *SE* |
| 1 | .004 | .080 | 1,139 | 61 | 1,200 | .044 | .058 | .031 | .047 |
| 2 | .080 | .122 | 1,097 | 103 | 1,200 | .003 | .045 | .000 | .037 |
| 3 | .122 | .196 | 1,044 | 156 | 1,200 | .075 | .041 | .016 | .034 |
| 4 | .196 | .327 | 856 | 344 | 1,200 | −.057[a] | .028 | −.023 | .024 |
| 5 | .327 | .830 | 644 | 556 | 1,200 | −.034 | .029 | −.036 | .025 |
| 4a[†] | .196 | .265 | 446 | 154 | 600 | −.055 | .044 | −.002 | .038 |
| 4b[†] | .266 | .327 | 410 | 190 | 600 | −.054 | .037 | −.041 | .031 |
| 5a[††] | .327 | .363 | 184 | 116 | 300 | .009 | .059 | −.031 | .050 |
| 5b[††] | .363 | .417 | 166 | 134 | 300 | −.039 | .062 | −.035 | .051 |
| 5c[††] | .417 | .512 | 158 | 142 | 300 | −.077 | .048 | −.064 | .043 |
| 5d[††] | .513 | .830 | 136 | 164 | 300 | −.060 | .061 | −.045 | .054 |

*Note.* ANCOVA = analysis of covariance.
[a] Significant at .05 level. [†] Division of Class 4. [††] Division of Class 5.

## Dual-Modeling Strategies

### *Method 7: Weighted Residual Bias Corrections*

Propensity scores are effective for removing bias, whereas models for potential outcomes help to increase efficiency. This suggests that the two may be combined to produce estimates with even better properties. Advantages of dual modeling were already seen in the results from matching and subclassification. Applying ANCOVA in a propensity-matched sample or within propensity-defined subclasses produced better estimates than matching or subclassification alone. In this section, we describe more ways to combine estimated propensity scores with models for the potential outcomes.

Consider the regression estimate for ACE (Equation 15), the average difference between the regression predictions for $Y_{i1}$ and $Y_{i0}$. If the model fit to treated persons does not accurately describe how $Y_{i1}$ varies with $X_i$, then the average of the $\hat{Y}_{i1}$s in the sample may be a biased estimate of $E(Y_{i1})$. Similarly, if the model fit to untreated persons does not accurately describe how $Y_{i10}$ varies with $X_i$, then the sample average of the $\hat{Y}_{i0}$s may be a biased estimate of $E(Y_{i0})$. It is difficult to guess the direction or magnitude of these biases with standard regression diagnostics. If propensity scores are available, however, estimates of these biases can be constructed through a clever combination of propensities and residuals.

Let $\hat{\epsilon}_{i1} = Y_{i1} - \hat{Y}_{i1}$ be the residual for person $i$ in the regression of $Y_{i1}$ on $X_i$. This residual is seen for every treated person. The average value of these residuals in the treated sample is essentially zero. (If the model is fit by OLS, it will be exactly zero.) The average residual in the full sample, however, could be positive or negative, because if the regression relationships have not been correctly specified, the model may tend to overpredict or underpredict $Y_{i1}$. We cannot compute the average of $\hat{\epsilon}_{i1}$ in the full sample, but we can estimate it using inverse-propensity weights. The weighted average $\sum_i T_i \hat{\pi}_i^{-1} \hat{\epsilon}_{i1} / \sum_i T_i \hat{\pi}_i^{-1}$ is an approximately unbiased estimate of this average. It is also an estimate of (minus one times) the bias in $\sum_i \hat{Y}_{i1}/N$ the regression estimate of $E(Y_{i1})$. Adding this weighted average to the regression estimate produces a bias-corrected estimate of $E(Y_{i1})$. In a similar fashion, residuals from untreated persons may be used to correct bias in $\sum_i \hat{Y}_{i1}/N$ the regression estimate of $E(Y_{i0})$. The correction term for that estimate is the weighted average of the residuals $\hat{\epsilon}_{i0} = Y_{i0} - \hat{Y}_{i0}$ in the untreated sample, using weights $1/(1 - \hat{\pi}_i)$. The difference between the corrected estimates of $E(Y_{i1})$ and $E(Y_{i0})$, which can be written as

$$\hat{ACE} = \frac{1}{N}\sum_i (\hat{Y}_{i1} - \hat{Y}_{i0}) + \frac{\sum_i T_i \hat{\pi}_i^{-1} \hat{\epsilon}_{i1}}{\sum_i T_i \hat{\pi}_i^{-1}} - \frac{\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1} \hat{\epsilon}_{i0}}{\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1}}, \quad (22)$$

is a bias-corrected estimate of ACE. On the right-hand side of Equation 22, the first term is the uncorrected regression estimate of $E(Y_{i1} - Y_{i0})$; the second term is the bias correction for $E(Y_{i1})$; and the third term is the bias correction for $-E(Y_{i0})$.

By the same principle, we can repair the bias in a regression estimate of $ACE_1$. The estimate of Equation 17 uses only one regression, a model for $Y_{i0}$, fit to the untreated sample. A weighted average of the residuals $\hat{\epsilon}_{i0}$ from that sample, using the weights $\hat{\pi}_i/(1 - \hat{\pi}_i)$, estimates the aver-

age of the residuals among treated persons. Subtracting this weighted average from the estimate in Equation 17 gives a bias-corrected estimate of $ACE_1$,

$$\hat{ACE}_1 = \frac{\sum_i T_i(Y_i - \hat{Y}_{i0})}{\sum_i T_i} - \frac{\sum_i (1 - T_i)\hat{\pi}_i(1 - \hat{\pi}_i)^{-1}\hat{\epsilon}_{i0}}{\sum_i (1 - T_i)\hat{\pi}_i(1 - \hat{\pi}_i)^{-1}}. \quad (23)$$

Standard errors for the estimates in Equations 22–23 are derived in the Appendix for the case where $\hat{Y}_{i0}$ and $\hat{Y}_{i1}$ are obtained from OLS and $\hat{\pi}_i$ comes from logistic regression.

Regression estimates with weighted residual bias corrections were first proposed by Cassel, Särndal, and Wretman (1976, 1977) as the basis for a methodology called model-assisted survey estimation (Särndal, Swensson, & Wretman, 1992). Model-assisted survey estimates use regression to predict the responses for unsampled persons, but they incorporate corrections to wipe out biases that may arise if the regression model is wrong. A similar strategy was independently proposed by Robins et al. (1994, 1995), but in a very different context, and with a different motivation. Robins et al. described methods for estimating regression coefficients from incomplete data. They first eliminated selection bias by IPW, weighting each observation by the inverse of the estimated probability of not being missing. Realizing that IPW can be inefficient, they enlarged the class of IPW estimates to incorporate regression predictions for the missing values. Theoretical large-sample properties of these methods were explored by Robins and Rotnitzky (1995), who concluded that an estimate with a form similar to Equation 22 was the most efficient one within the class. We presented this method as a bias-corrected version of regression estimation. Robins et al. motivated the method as an efficiency-enhanced version of IPW. Either way, it represents an ingenious combination of regression predictions and propensity scores.

We have argued that the estimates in Equations 22–23 are consistent if the regressions for $Y_{i1}$ and $Y_{i0}$ are wrong, as long as the model used to estimate the $\pi_i$s is correct. The reverse is also true. If models for $Y_{i1}$ and $Y_{i0}$ are correct, then the residuals have mean zero in the population and in every subpopulation defined by $X_i$, and the correction terms vanish on average, *whether the propensity model is right*. This is known as double robustness (van der Laan & Robins, 2003). Doubly robust estimators are discussed by Bang and Robins (2005); Carpenter, Kenward, and Vansteelandt (2006); Davidian, Tsiatis, and Leon (2005); and Lunceford and Davidian (2004). Double robustness is theoretically interesting, but what matters in practice is how these estimators perform when none of the models are precisely true (Kang & Schafer, 2007).

To apply this method to our simulated observational study, we fit a logistic model to $T_i$ and linear regressions to $Y_{i1}$ and $Y_{i0}$ with main effects for all 13 baseline measures. In our first sample, the estimate of ACE was .006 ($SE$ = .022), and the estimate of $ACE_1$ was −.020 ($SE$ = .015). If the method works as advertised, it should be less biased than regression estimation and more efficient than IPW. Over the 5,000 samples, it was indeed more efficient than IPW for estimating ACE. The RMSE dropped from .027 to .023, a decrease of about 15%. However, the biases in the two procedures were about the same. The bias in the IPW estimate, which arises because the logistic model does not accurately predict propensities near zero, has been carried over to the new method. Comparing the new method with regression estimation, however, we find that the bias in the estimate of ACE (which was already small) has actually *increased*. Confidence intervals for ACE from the regression estimates had a coverage rate of 94.3%, but the coverage rate for the new method is only 86.2%. The regression models were not perfectly unbiased, but our attempt to correct their biases through an imperfect propensity model actually made matters worse.

This example demonstrates that the clever bias-corrected estimates of Equations 22–23 are sometimes inferior to uncorrected regression estimates. The alluring property of double robustness may not translate into actual robustness when both models are incorrect. Weighted residual bias corrections are susceptible to the same problems as ordinary IPW estimation when a few individuals are given large weights. Regression estimates may be biased, but weighted residual bias corrections may not be the best way to repair them.

### *Method 8: Weighted Regression Estimation*

Another way to correct a regression estimate for bias is to apply inverse-probability weights while fitting the regression models. In typical regression settings, weights are used to correct for heteroscedasticity, to improve the efficiency of estimates, and to improve the accuracy of standard errors. In this context, however, the weights assume a different role: to give consistent estimates of the regression coefficients that would result from fitting these models to all potential outcomes in the population.

As before, suppose we apply linear regression models to the potential outcomes, $E(Y_{i1}|X_i) = X_i^T\beta_1$ and $E(Y_{i0}|X_i) = X_i^T\beta_0$, where the vector of covariates, $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$, includes a constant term ($X_{i1} = 1$). Suppose that the vectors of regression coefficients $\beta_1$ and $\beta_0$ are now estimated by WLS, using weights $1/\hat{\pi}_i$ and $1/(1 - \hat{\pi}_i)$.

$$\hat{\beta}_1 = \left(\sum_i T_i\hat{\pi}_i^{-1}X_iX_i^T\right)^{-1}\left(\sum_i T_i\hat{\pi}_i^{-1}X_iY_i\right) \quad (24)$$

for the treated sample, and

$$\hat{\beta}_0 = \left( \sum_i (1 - T_i)(1 - \hat{\pi})^{-1} X_i X_i^T \right)^{-1} \times \left( \sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1} X_i Y_i \right) \quad (25)$$

for the untreated sample. Let $\hat{Y}_{i1} = X_i^T \hat{\beta}_1$ and $\hat{Y}_{i0} = X_i^T \hat{\beta}_0$ denote the predicted values of $Y_{i1}$ and $Y_{i0}$, which are computed for the full sample. The weighted regression estimate for $\hat{ACE} = \sum_i (\hat{Y}_{i1} - \hat{Y}_{i0})/N$. For estimating $\text{ACE}_1$, we would omit the model for $Y_{i1}$ and use weights $\hat{\pi}_i/(1 - \hat{\pi}_i)$ instead of $1/(1 - \hat{\pi}_i)$ in the estimation of $\beta_0$. The estimate for $\text{ACE}_1$ would then be $\hat{ACE}_1 = \sum_i T_i (Y_{i1} - \hat{Y}_{i0})/\sum_i T_i$. These weighted regression estimates are doubly robust (Kang & Schafer, 2007). Standard errors for these estimates are derived in the Appendix.

In our first sample from the artificial population, the weighted regression method gave $\hat{ACE} = .000$ ($SE = .018$) and $\hat{ACE}_1 = -.020$ ($SE = .015$). Over repeated samples, estimates for ACE were slightly less biased and more efficient than those from the weighted residual method. However, the bias was still large enough to impair intervals; coverage for ACE was 88.4%. Like any method that uses inverse-propensity weighting, this procedure is susceptible to bias from a misspecified propensity model when some $\hat{\pi}_i$s are close to zero or one.

### *Method 9: Regression Estimation With Propensity-Related Covariates*

The weighted residual method applies a post hoc correction to a regression estimate from a faulty model. The weighted regression method adjusts the estimated coefficients from the faulty model. If the model is faulty, however, then perhaps it would be better to improve the model by changing the assumed form of $E(Y_{i1}|X_i)$ or $E(Y_{i0}|X_i)$.

If a regression model incorrectly characterizes relationships between the outcome and the predictors, the problem can be diagnosed by plotting the residuals versus the predicted values, or versus individual predictors. Nonlinear trends (e.g., curvature) in any of these plots indicate that the predictions are too high for some types of individuals and too low for others. If the model is correctly specified, there should be no tendency for residuals to rise or fall in relation to any predictor or any combination of predictors.

To obtain an unbiased regression estimate of an ACE, we do not need our models for $Y_{i1}$ and $Y_{i0}$ to give unbiased predictions for all values of every covariate. From the balancing property of the propensity score (Equation 19), it can be shown that we only need unbiased prediction within classes of constant propensity. We may diagnose the latter by plotting residuals against the estimated propensity scores or against the logit-propensity scores.

To illustrate, we fit linear regressions to $Y_{i1}$ and $Y_{i0}$ using the dieters and nondieters in our initial sample. As before, each model included all 13 baseline variables. We also estimated the propensities to diet by logistic regression using the same 13 predictors. Plots of the residuals $\hat{\epsilon}_{i0} = Y_{i0} - \hat{Y}_{i0}$ and $\hat{\epsilon}_{i1} = Y_{i1} - \hat{Y}_{i1}$ from the linear models versus estimated logit-propensities for nondieters and dieters are shown in Figure 4. Because the logit-propensity is a linear combination of predictors in the models, there should be no relationship between the logit-propensities and residuals in either plot. Indeed, linear relationships estimated by OLS, represented by the dashed lines, have intercepts and slopes of zero. Smooth trends estimated by local polynomial

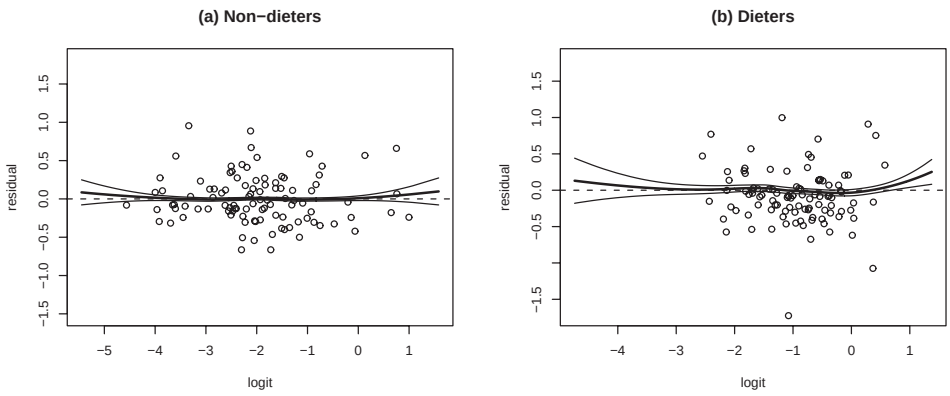


*Figure 4.* Plots of residuals from linear regressions of emotional distress at Wave II on baseline variables, versus estimated logit-propensities to diet, for (a) nondieters and (b) dieters in one sample of $N = 6{,}000$ girls from the synthetic population. (To avoid overplotting, only 100 randomly selected points are shown in each plot.) Dashed straight lines show linear fits. The solid dark curves indicate the nonparametric loess fits; the solid light curves indicate pointwise 95% confidence intervals around the loess fits.

regression (loess; Cleveland & Devlin, 1988; Cohen et al., 2003, chapter 4), however, tell a different story. In each plot, the loess-estimated means of the residuals for those with the lowest and highest propensities to diet are positive, suggesting that the predictions of $Y_{i0}$ and $Y_{i1}$ for these individuals may be too low. For dieters, a slight elevation in the loess fit in the middle range of propensities suggests that their predictions for $Y_{i1}$ may also be too low. These plots suggest that the fit of our models with respect to the propensity score can be improved.

Little and An (2004) created a regression estimate from a model that allowed the mean response to vary with propensities in a flexible way, describing the relationship by a cubic spline. On the basis of our previous work (Kang & Schafer, 2007), we suggest a simpler method: (a) Classify individuals into at $S \geq 5$ classes in which the estimated propensity scores are nearly homogeneous; (b) create $S - 1$ dummy variables to distinguish among the classes; (c) include these dummy variables as predictors in the regressions for $Y_{i1}$ and $Y_{i0}$; and (d) compute regression estimates for ACE or $ACE_1$ in the usual way (Equation 15 or 17). Following Cochran (1968), the use of five classes should remove about 90% of the bias in a regression estimate due to misspecification of the regression relationships.

Another way to use propensities as covariates was proposed by Bang and Robins (2005). They included $1/\hat{\pi}_i$ as a predictor for $Y_{i1}$, and $1/(1 - \hat{\pi}_i)$ as a predictor for $Y_{i0}$, and computed a regression estimate for ACE as in Equation 15. Their method has been shown to be doubly robust (Scharfstein, Rotnitzky, & Robins, 1999), and they recommended that it be used routinely when estimating treatment effects from observational data. We *do not* endorse this method. It can misbehave when the potential-outcome and propensity models are both incorrect, and it is very sensitive to misspecification of the propensity model where $\hat{\pi}_i \approx 0$ and $\hat{\pi}_i \approx 1$ (Kang & Schafer, 2007).

Using our initial sample of $N = 6{,}000$, we fit a logistic propensity model and classified individuals into the nine classes shown in Table 5. We created eight dummy variables to distinguish among the classes, included these dummy variables in the linear regression models for $Y_{i1}$ and $Y_{i0}$, and computed regression estimates in the usual way. The resulting estimates were $\hat{ACE} = -.001$ ($SE = .016$) and $\hat{ACE}_1 = -.020$ ($SE = .014$). Over 5,000 repeated samples, this method performed extremely well. It had the smallest bias of any method we tried for estimating ACE, was highly efficient, and produced intervals with excellent coverage.

We have frequently been asked whether it would be appropriate to include the propensity score itself as a covariate in ANCOVA. If the propensities are estimated by logistic regression from covariates in the ANCOVA model, then the estimated logit-propensity will be a linear combination of predictors already in the model. Without a logit transformation, the propensity will not be perfectly collinear with the other variables. However, we do not recommend using it as a covariate, as this would assume it is linearly related to each potential outcome. In our experience, the biases suggested by residual plots (e.g., as shown in Figure 4) can rarely be corrected by a linear trend in the propensity score, or even by a quadratic or cubic trend. Dummy variables or splines tend to perform much better.

## Discussion

### *Comparing the Performance of the Methods*

Rubin's causal model clarifies the meaning of an ACE, but we have a great variety of methods for estimating ACEs. Estimates and standard errors from all the methods that we applied to our initial sample are reported in Table 6. A standard error from one sample measures the degree to which the estimate will vary over repeated samples. In large observational studies, however, the most harmful component of error in an estimate is not sampling variation but bias arising from failure of the underlying assumptions. This kind of bias does not vanish as the sample size grows. In fact, it becomes more worrisome because, as $N \rightarrow \infty$, standard errors shrink to a point where even a small bias can seriously impair the performance of intervals and tests.

The performance of the methods over 5,000 samples is summarized in Table 7. The columns of Table 7 report the bias in the estimates, standard deviation of the estimates, RMSE, percentage of nominal 95% intervals covering the true population values, and average width of the intervals. Examining the RMSE, we find that the most precise estimates of ACE came from ANCOVA with Baseline $\times$ Treatment interactions (and the equivalent regression estimates), subclassification on propensity scores with ANCOVA, and regression estimation with propensity-related covariates. Each of these had low bias and good coverage. The worst estimates came from the difference in means, followed by IPW, regression estimation with weighted residual bias corrections, and linear ANCOVA without interactions. Procedures that relied on inverse-propensity weights—IPW, weighted regression, and weighted residual bias corrections—did not perform well for ACE, because only a few dieters had estimated propensities near zero. Weights for estimating the population mean of $Y_{i1}$ were heavily concentrated on these individuals. When a few weights become very large, weighting is inefficient. It also becomes susceptible to bias, because misspecification of the propensity model can produce large errors in the most extreme weights.

For estimating $ACE_1$, all of the methods performed well except ANCOVA without interactions and matching followed by a *t*-test. Conditions for estimating $ACE_1$ were favorable in this example, because the propensity score distribution for the treated was well covered by the distribution for the untreated. For any girl who dieted, it was easy

Table 6
*Estimates and Standard Errors for Average Causal Effect (ACE) of Dieting on Emotional Distress in a Population of Adolescent Girls (ACE) and Among Girls Who Dieted ($ACE_1$), Based on a Simulated Sample of N = 6,000 Girls*

| Method | $ACE = .003$ | | $ACE_1 = -.022$ | |
|---|---|---|---|---|
| | Estimate | *SE* | Estimate | *SE* |
| 1. Difference in means (*t*-test) | .060 | .015 | | |
| 2. ANCOVA (main effects only) | −.014 | .013 | −.014 | .013 |
| ANCOVA (with interactions) | −.006 | .015 | −.015 | .013 |
| 3. Regression estimation | −.006 | .016 | −.015 | .014 |
| 4. Matching + *t*-test | | | .009 | .019 |
| Matching + ANCOVA | | | −.018 | .016 |
| 5. Inverse-propensity weighting | −.005 | .028 | −.025 | .015 |
| 6. Subclassification + *t*-test | .005 | .019 | −.024 | .017 |
| Subclassification + ANCOVA | −.004 | .015 | −.023 | .015 |
| 7. Weighted residual bias correction | .006 | .022 | −.020 | .015 |
| 8. Weighted regression estimation | −.000 | .018 | −.020 | .015 |
| 9. Propensity-related covariates | −.001 | .016 | −.020 | .014 |

*Note.* ANCOVA = analysis of covariance.

to find girls with similar propensities who did not, so the data contained ample information to predict the missing $Y_{i0}$s. Moreover, none of the propensities were close to one, so weighting methods that used $(1 - \hat{\pi}_i)$ in the denominator remained stable. Under conditions like these, many different methods will give reasonable estimates of $ACE_1$.

No simulation can represent all situations encountered in practice. Ours resembles an observational study with a large sample size and pre–post measurements on the outcome that are moderately highly correlated ($r \approx .6$). In earlier simulations with $N = 1{,}000$, we found that biases were about the same as in Table 7, but standard deviations were approximately multiplied by $\sqrt{6} \approx 2.5$. Under those conditions, intervals had better coverage for all methods. In psychological studies with smaller *N*, we would expect the differences between methods to diminish as bias becomes a relatively less important part of the overall uncertainty. Moreover, in situations where relationships between the covariates and outcomes are stronger, the impact of bias can also become stronger. In an earlier version of this article, we constructed a simulation to mimic a study of the effect of dieting on body mass index (BMI). The correlation between outcome BMI and baseline BMI was $r \approx .9$, and regression-based methods became more powerful and more susceptible to bias resulting from model failure (e.g., nonlinearity).

Table 7
*Performance of Procedures for Estimating Average Causal Effect (ACE) of Dieting on Emotional Distress in a Population of Adolescent Girls (ACE) and Among Girls Who Dieted ($ACE_1$) Over 5,000 Simulated Samples of N = 6,000: Bias, Standard Deviation of Estimates, Root-Mean-Square Error (RMSE), Percent Coverage of Nominal 95% Confidence Intervals, and Average Interval Width*

| Method | $ACE = .003$ | | | | | $ACE_1 = -.022$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Bias | *SD* | RMSE | Coverage | Width | Bias | *SD* | RMSE | Coverage | Width |
| 1. Difference in means (*t*-test) | .052 | .016 | .054 | 8.7 | .120 | | | | | |
| 2. ANCOVA (main effects only) | −.016 | .014 | .021 | 76.7 | .104 | .009 | .014 | .016 | 89.0 | .104 |
| ANCOVA (with interactions) | −.004 | .015 | .016 | 94.2 | .122 | .006 | .014 | .015 | 91.6 | .106 |
| 3. Regression estimation | −.004 | .015 | .016 | 94.3 | .122 | .006 | .014 | .015 | 93.2 | .111 |
| 4. Matching + *t*-test | | | | | | .035 | .017 | .039 | 58.6 | .155 |
| Matching + ANCOVA | | | | | | .004 | .016 | .017 | 94.3 | .129 |
| 5. Inverse-propensity weighting | −.014 | .022 | .027 | 88.8 | .176 | .002 | .016 | .016 | 94.8 | .122 |
| 6. Subclassification + *t*-test | .005 | .018 | .018 | 96.1 | .152 | .008 | .015 | .017 | 95.8 | .136 |
| Subclassification + ANCOVA | .001 | .016 | .016 | 94.4 | .125 | .006 | .015 | .016 | 93.6 | .117 |
| 7. Weighted residual bias correction | −.013 | .018 | .023 | 86.2 | .142 | .005 | .015 | .016 | 93.7 | .120 |
| 8. Weighted regression estimation | −.009 | .016 | .019 | 88.4 | .123 | .005 | .015 | .016 | 93.7 | .118 |
| 9. Propensity-related covariates | .000 | .016 | .016 | 94.6 | .127 | .005 | .015 | .016 | 93.5 | .115 |

*Note.* ANCOVA = analysis of covariance.

### *Lessons Learned*

In this article, we have reviewed part of the vast and growing literature on causal inference. Through our review and simulated study, we learned several lessons that seem generally relevant to observational studies in psychology.

One important lesson is that estimates from ANCOVA based on the simple linear model (Equation 6) are often untrustworthy. However, the performance of ANCOVA can be greatly enhanced by taking the following steps.

1. Consider the possibility of nonlinear trends with respect to important baseline covariates (Little et al., 2000).
2. Include summaries of the propensity scores (e.g., four dummy variables to distinguish among five classes) as additional baseline variables.
3. Include Baseline $\times$ Treatment interactions.
4. Express the results as regression estimates and compute robust standard errors by the procedures given in the Appendix.

With these enhancements, ANCOVA is no longer one of the worst ways to estimate an ACE, and it may be one of the best.

A second lesson we have learned is that dual modeling of the propensity scores and potential outcomes is a commendable idea, but it matters how information from the two sources is combined. Fitting ANCOVA models to propensity-matched samples, or within classes defined by propensity scores, is usually better than matching or subclassification alone. Including summaries of the propensity scores as additional covariates may improve the performance of ANCOVA. However, using propensities in a weighted residual bias correction, or in a weighted regression estimate, is sometimes unwise and can make matters worse.

The third lesson pertains to the central role of the propensity score. Propensities are invaluable for diagnosing overlap and imbalance. Modeling the treatment indicator and examining the distributions of the logit-propensities should be one of the very first steps taken by an analyst. Propensity scores are also helpful in matching, subclassification, and defining covariates for ANCOVA models. However, we do not advocate using reciprocals of propensities as weights. Weighting sometimes works well, but it often fails. At best, it performs only a little better than matching, subclassification, or propensity-defined covariates. At worst, it is much less efficient and very sensitive to misspecification of the propensity model. Other methodologists may disagree, but we have not yet found any compelling theoretical or practical reason to use propensities in weights.

We have also learned the value of applying multiple methods to estimate ACEs. In real examples, it is always advisable to try a variety of methods. Results from these methods may not agree, and making sense of the differences inevitably alerts us to interesting and important features of the data that we would not have otherwise noticed. Any of the methods we have described may give a reasonable estimate and standard error for ACE if the assumptions underlying the method are satisfied. When answers differ, we need to examine the data more closely to look for departures from the assumed models and to correct for deficiencies. In our first simulated sample of $N = 6,000$, for example, we noted that the IPW estimate of the population ACE had a much larger standard error than the ANCOVA estimates did. This fact immediately led us to examine the distribution of the weights $1/\hat{\pi}_i$, which revealed that IPW gave undue weight to a small number of girls with very low propensities to diet.

Finally, we discourage researchers from choosing a methodology solely on its theoretical mathematical properties, because these may assume conditions that are not satisfied in practice. The arguments of Robins and Rotnitzky (1995), for example, suggest that a regression estimate with weighted residual bias correction should be nearly impossible to beat. In our simulations, however, that method gave poor results for ACE. In our opinion, a method that maintains good performance when both models are moderately misspecified is preferable to one that is highly optimized for an ideal situation where both models are true.

### *Topics Not Covered*

In this article, we have not discussed the use of instrumental variables. Those methods, which have a long history in economics, require an instrument, a variable that is believed to affect the outcome only through the treatment. We omitted this topic because in psychological research the presence of an instrument tends to be the exception rather than the rule. Instrumental variables and their relationship to potential outcomes are discussed by Angrist, Imbens, and Rubin (1996); Gelman and Hill (2007); and Winship and Morgan (1999).

Space limitations have precluded us from discussing multiple imputation (MI) for potential outcomes. MI is an attractive, practical solution to many missing-data problems (Rubin, 1987; Schafer & Graham, 2002) and may be viewed as a computational device for approximating a fully parametric Bayesian analysis (Dominici, Zeger, Parmigiani, Katz, & Christian, 2006). In causal inference, MI would require us to make assumptions about the inestimable partial correlation between $Y_{i1}$ and $Y_{i0}$ given the covariates. Although this may seem troubling, inferences about ACEs are insensitive to what we assume about this parameter (Gelman, Carlin, Stern, & Rubin, 2004).

## References


Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88,* 669–679.

Allison, P. D. (2001). *Missing data.* Thousand Oaks, CA: Sage.

Angrist, J. D., Imbens, G., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91,* 444–455.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics, 61,* 962–972.

Berman, N. C., & Greenhouse, S. W. (1992). Comment: Adjusting for demographic covariates by the analysis of covariance [with rejoinder]. *Journal of Clinical and Experimental Neuropsychology, 14,* 981–985.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. M. (1984). *Classification and regression trees.* Boca Raton, FL: CRC Press.

Carpenter, J., Kenward, M., & Vansteelandt, S. (2006). A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal Statistical Society, Series A, 169,* 571–584.

Cassel, C. M., Särndal, C. E., & Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika, 63,* 615–620.

Cassel, C. M., Särndal, C. E., & Wretman, J. H. (1977). *Foundations of inference in survey sampling.* New York: Wiley.

Chambers, R. L., & Skinner, C. J. (Eds.). (2003). *Analysis of survey data.* New York: Wiley.

Cleveland, W. S., & Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association, 83,* 596–610.

Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A, 128,* 234–265.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24,* 205–213.

Cochran, W. G. (1977). *Sampling techniques* (2nd ed.). New York: Wiley.

Cochran, W. G. (1983). *Planning and analysis of observational studies.* New York: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive missing-data strategies in modern missing-data procedures. *Psychological Methods, 6,* 330–351.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton-Mifflin.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., & Wynder, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute, 22,* 173–203.

D'Agostino, R. B., Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17,* 2265–2281.

Davidian, M., Tsiatis, A. A., & Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest-posttest study without missing data. *Statistical Science, 20,* 261–301.

Dawid, A. P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association, 95,* 407–448.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94,* 1053–1062.

DeMaris, A. (2002). Explained variance in logistic regression. *Sociological Methods and Research, 31,* 27–74.

Dominici, F., Zeger, S. L., Parmigiani, G., Katz, J., & Christian, P. (2006). Estimating percentile-specific treatment effects in counterfactual models: A case-study of micronutrient supplementation, birth weight and infant mortality. *Applied Statistics, 55,* 261–280.

DuMouchel, W. H., & Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association, 78,* 535–543.

Elliott, M. R., & Little, R. J. A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics, 16,* 191–209.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman and Hall.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

Gelman, A., & Meng, X. L. (Eds.). (2004). *Applied Bayesian modeling and causal inference from incomplete-data perspectives.* New York: Wiley.

Gerner, B., & Wilson, P. H. (2005). The relationship between friendship factors and adolescent girls' body image concern, body dissatisfaction, and restrained eating. *International Journal of Eating Disorders, 37,* 313–320.

Green, M. W., & Rogers, P. J. (1995). Impaired cognitive function during spontaneous dieting. *Psychological Medicine, 25,* 1003–1010.

Greenland, S., Pearl, J., & Robins, J. (1999). Causal diagrams for epidemiologic research. *Epidemiology, 10,* 37–48.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica, 15,* 413–419.

Hardin, J. W., & Hilbe, J. M. (2003). *Generalized linear models.* New York: Chapman & Hall/CRC Press.

Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods, 12,* 247–267.

Heatherton, T. F., & Polivy, J. (1992). Chronic dieting and eating disorders: A spiral model. In J. H. Crowther, D. L. Tennenbaum, S. E. Hobfoll, & M. A. P. Stephens (Eds.), *The etiology of bulimia nervosa: The individual and family context* (pp. 133–148). London: Hemisphere Publishing.

Heckman, J. J., Smith, J., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies, 64,* 487–535.

Hernan, M., Brumback, B., & Robins, J. M. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11*(5), 550–560.

Hill, A. J. (2004). Does dieting make you fat? *British Journal of Nutrition, 92*(Suppl. 1), S15–S18.

Hinkley, D. (1985). Transformation diagnostics for linear models. *Biometrika, 72,* 487–496.

Hirano, K., & Imbens, G. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catherization. *Health Services and Outcome Research Methodology, 2,* 259–278.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2004). *MatchIt: Nonparametric preprocessing for parametric causal inference.* Statistical software package for R, available from http://www.gking.harvard.edu

Höfler, M. (2005). Causal inference based on counterfactuals. *BMC Medical Research Methodology, 5,* 28.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–970.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 27,* 205–224.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47,* 663–685.

Hosmer, D. W., & Lemeshow, S. (2003). *Applied logistic regression* (2nd ed.). New York: Wiley.

Hsu, L. K. (1996). Epidemiology of the eating disorders. *Psychiatric Clinics of North America, 19,* 681–700.

Huitema, B. (1980). *Analysis of covariance and alternatives.* New York: Wiley.

Jaccard, J., & Wan, C. K. (1996). *LISREL approaches to interaction effects in multiple regression.* Thousand Oaks, CA: Sage.

Johnson, F., & Wardle, J. (2005). Dietary restraint, body dissatisfaction, and psychological distress: A prospective analysis. *Journal of Abnormal Psychology, 115,* 119–125.

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating population means from incomplete data. *Statistical Science, 26,* 523–539.

Katz, D. L. (2005). Competing dietary claims for weight loss: Finding the forest through truculent trees. *Annual Review of Public Health, 26,* 61–88.

King, G. (2006). The dangers of extreme counterfactuals. *Political Analysis, 14,* 131–159.

King, G., & Zeng, L. (2002). Improving forecasts of state failure. *World Politics, 53,* 623–658.

King, G., & Zeng, L. (2006). When can history be our guide? The pitfalls of counterfactual inference. *International Studies Quarterly, 51,* 183–210.

Kovacs, M., Obrosky, D. S., & Sherrill, J. (2003). Developmental changes in the phenomenology of depression in girls compared to boys from childhood onward. *Journal of Affective Disorders, 74,* 33–48.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (4th ed.). New York: McGraw-Hill.

Lalonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review, 76,* 604–620.

Leuven, E., & Sianesi, B. (2003). *PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.* Statistical Software Components S432001, Boston College Department of Economics, Boston College.

Levine, M. P., Smolak, L., & Hayden, H. (1994). The relation of sociocultural factors to eating attitudes and behaviors among middle school girls. *Journal of Early Adolescence, 14,* 471–490.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13–22.

Little, R. J. A., & An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica, 14,* 949–968.

Little, R. J. A., An, H., Johanns, J., & Giordani, B. (2000). A comparison of subset selection and analysis of covariance for the adjustment of confounders. *Psychological Methods, 5,* 459–476.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Liu, C. (2004). Robit regression: A simple robust alternative to logistic and probit regression. In A. Gelman & X. L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 227–238). New York: Wiley.

Lohr, S. (1999). *Sampling: Design and analysis.* Pacific Grove, CA: Duxbury Press.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29,* 530–558.

Lunceford, J. K., & 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.

Manski, C. F. (1999). Comment: Choice as an alternative to control in observational studies. *Statistical Science, 14,* 279–281.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9,* 403–425.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110,* 40–48.

Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics, 56,* 118–124.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research.* New York: Cambridge University Press.

Neumark-Sztainer, D., & Hannan, P. J. (2000). Weight-related behaviors among adolescent girls and boys: Results from a national survey. *Archives of Pediatrics and Adolescent Medicine, 154,* 569–577.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. In Z. Griliches (Ed.), *Handbook of econometrics* (Vol. 4, pp. 2111–2245). Amsterdam: Elsevier.

Neyman, J. (1990). On the application of probability theory to agricultural experiments: Essays on Principles, Section 9. *Translated in Statistical Science, 5,* 465–480. (Original work published 1923)

Parsons, L. J. (2000). Using SAS software to perform a case-control match on propensity score in an observational study. *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference,* Paper 225–25, Cary, NC: SAS Institute.

Patton, G. C., Selzer, R., Coffey, C., Carlin, J. B., & Wolfe, R. (1999). Onset of adolescent eating disorders: Population based cohort study over 3 years. *British Medical Journal, 318,* 765–768.

Pearl, J. (2000). *Causality: Models, reasoning and inference.* Cambridge, England: Cambridge University Press.

Pfefferman, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61,* 317–337.

Pfefferman, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research, 5,* 239–261.

Pratt, J. W., & Schlaifer, R. (1988). On the interpretation and observation of laws. *Journal of Econometrics, 39,* 23–52.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics, 38,* 485–498.

Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., et al. (1997). Protecting adolescents from harm: Findings from the National Longitudinal Study on Adolescent Health. *Journal of the American Medical Association, 278,* 823–832.

Robins, J. (1999). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology: The environment and clinical trials* (pp. 95–134). New York: Springer-Verlag.

Robins, J. M., Hernan, M., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11,* 550–560.

Robins, J. M., & 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., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association, 90,* 106–121.

Rosen, J. C., Tacy, B., & Howell, D. (1990). Life stress, psychological symptoms and weight reducing behavior in adolescent girls: A prospective analysis. *International Journal of Eating Disorders, 9,* 17–26.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, 147,* 656–666.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using sub-classification on the propensity score. *Journal of the American Statistical Association, 79,* 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39,* 33–38.

Rotnitzky, A., Robins, J. M., & Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with ignorable nonresponse. *Journal of the American Statistical Association, 93,* 1321–1339.

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics, 29,* 185–203.

Rubin, D. B. (1974a). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association, 69,* 467–474.

Rubin, D. B. (1974b). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6,* 34–58.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74,* 318–328.

Rubin, D. B. (1980). Discussion of "randomization analysis of experimental data in the Fisher randomization test" by D. Basu. *Journal of the American Statistical Association, 75,* 591–593.

Rubin, D. B. (1984). William G. Cochran's contributions to the design, analysis and evaluation of observational studies. In P. S. R. S. Rao & J. Sedransk (Eds.), *W. G. Cochran's impact on statistics* (pp. 37–69). New York: Wiley.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, D. B. (1990). Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science, 5,* 472–480.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127,* 757–763.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2,* 169–188.

Rubin, D. B. (2004). Causal effects via potential outcomes. *Scandinavian Journal of Statistics, 31,* 161–170.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100,* 322–331.

Rubin, D. B. (2006). *Matched sampling for causal effects.* New York: Cambridge University Press.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52,* 249–264.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95,* 573–585.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling.* New York: Springer.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Schafer, J. L., & Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association, 95,* 144–154.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable dropout using semiparametric nonresponse models. *Journal of the American Statistical Association, 94,* 1096–1120 (with Rejoinder, 1135–1146).

Shadish, W. R., Clark, M. H., & Steiner, P. M. (in press). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association.*

Sobel, M. E. (1995). Causal inference in the social and behavioral sciences. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 1–38). New York: Plenum.

Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. *Methodology, 1,* 39–54.

Stice, E., Mazotti, L., Krebs, M., & Martin, S. (1998). Predictors of adolescent dieting behaviors: A longitudinal study. *Psychology of Addictive Behaviors, 12,* 195–205.

Stice, E., Shaw, H., & Nemeroff, C. (1998). Dual pathway model of bulimia nervosa: Longitudinal support for dietary restraint and affect-regulation mechanisms. *Journal of Social and Clinical Psychology, 17,* 129–149.

Udry, J. R. (2003). *The National Longitudinal Study of Adolescent Health (Add Health), Waves I and II, 1994–1996; Wave III, 2001–2002* (machine-readable data file and documentation). Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

van der Laan, M. J., & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality.* New York: Springer.

Vartivarian, S., & Little, R. J. A. (2002). On the formation of weighting adjustment cells for unit nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association,* 3553–3558.

Warren, C., & Cooper, P. J. (1988). Psychological effects of dieting. *British Journal of Clinical Psychology, 27,* 269–270.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–84). New York: Cambridge University Press.

White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica, 48,* 817–838.

Wilson, G. T. (1993). Relation of dieting and voluntary weight loss to psychological functioning and binge eating. *Annals of Internal Medicine, 119,* 727–730.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25,* 650–707.

Winship, C., & Sobel, M. E. (2004). Causal inference in sociological studies. In M. Hardy (Ed.), *Handbook of data analysis* (pp. 481–504). Thousand Oaks, CA: Sage.

Zanutto, E. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science, 4,* 67–91.

# Appendix

## Formulas for Variance Estimation

We now present formulas for computing standard errors for regression, inverse-propensity weighted (IPW), and dual-modeling estimates of average causal effect (ACE) and $ACE_1$. Our formulas are based on asymptotic arguments similar to those of Robins, Rotnitzky, and Zhao (1995). We assume that models for $Y_{i1}$ and $Y_{i0}$ are linear regressions fit by ordinary least squares (OLS) and that the model for $\pi_i$ is a logistic regression estimated by maximum likelihood.

For regression estimation, suppose that $E(Y_{it}|X_i) = X_i^T\beta_t$ for $t = 0, 1$. Define $\mu_0 = E(Y_{i0})$, and $\mu_1 = E(Y_{i1})$. Write ACE as $\mu_1 - \mu_0 = a^T\theta$, where $a = (0, \ldots, 0, -1, 1)^T$ *and* $\theta = (\beta_0^T, \beta_1^T, \mu_0, \mu_1)^T$. The estimate $\hat{ACE}$ in Equation 16 may be written as $a^T\hat{\theta}$, where $\hat{\theta}$ contains the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ (Equations 13-14), $\hat{\mu}_0 = N^{-1}\Sigma_i((1-T_i)Y_i + T_iX_i^T\hat{\beta}_0)$, and $\hat{\mu}_1 = N^{-1}\Sigma_i(T_iY_i + (1-T_i)X_i^T\hat{\beta}_1)$. We may regard $\hat{\theta}$ as the solution to a set of joint estimating equations $\Sigma_i\psi_i(\theta) = 0$, where the estimating functions are $\psi_i(\theta) = \psi_i = (S_{i0}^T, S_{i1}^T, U_{i0}, U_{i1})^T$, $S_{i0} = (1-T_i)(Y_i - X_i^T\beta_0)X_i$, $S_{i1} = T_i(Y_i - X_i^T\beta_1)X_i$, $U_{i0} = (1-T_i)(Y_{i-\mu 0}) + T_i(X_i\beta_0 - \mu_0)$, and $U_{i1} = T_i(Y_i - \mu_1) + (1-T_i)(X_i\beta_1 - \mu_1)$. By well-known properties of mean-zero estimating functions (e.g., Newey & McFadden, 1994), we can write $\sqrt{N}(\hat{\theta} - \theta) \rightarrow N(0,\Sigma)$, where $\Sigma = A^{-1}BA^{-1T}$, $A = -E(\delta\psi_i/\delta\theta^T)$, and $B = E(\psi_i\psi_i^T)$. The matrix $A$ can be partitioned as

$$A = \begin{bmatrix} A_{11} & 0 & 0 & 0 \\ 0 & A_{22} & 0 & 0 \\ A_{31} & 0 & A_{33} & 0 \\ 0 & A_{42} & 0 & A_{44} \end{bmatrix},$$

where $A_{11} = -E(\delta S_{i0}/\delta\beta_0^T)$, $A_{22} = -E(\delta S_{i1}/\delta\beta_1^T)$, $A_{31} = -E(\delta U_{i0}/\delta\beta_0^T)$, $A_{33} = -E(\delta U_{i0}/\delta\mu_0)$, $A_{42} = -E(\delta U_{i1}/\delta\beta_1^T)$, and $A_{44} = -E(\delta U_{i1}/\delta\mu_1)$. It follows that the variance of $\hat{ACE}$ may be approximated by $N^{-1}a^T\hat{A}^{-1}\hat{B}\hat{A}^{-1T}a$, where $\hat{B} = N^{-1}\Sigma_i\hat{\psi}_i\hat{\psi}_i^T$, $\hat{\psi}_i = \psi_i(\hat{\theta})$, and $\hat{A} = -N^{-1}\Sigma_i\delta\psi_i/\delta\theta^T$ evaluated at $\theta = \hat{\theta}$. The nonzero portions of $\hat{A}$ are $\hat{A}_{11} = N^{-1}\Sigma_i(1-T_i)X_iX_i^T$, $\hat{A}_{22} = N^{-1}\Sigma_iT_iX_iX_i^T$, $\hat{A}_{31} = -N^{-1}\Sigma_iT_iX_i^T$, $\hat{A}_{42} = -N^{-1}\Sigma_i(1-T_i)X_i^T$, and $\hat{A}_{33} = \hat{A}_{44} = 1$. Although we assume that $\beta_0$ and $\beta_1$ are estimated by OLS, this variance estimate is robust to departures from homoscedasticity in the models for the potential outcomes.

For the regression estimate of $ACE_1$ (Equation 17), the model for $Y_{i1}$ is eliminated. We now take $\hat{ACE}_1 = a^T\hat{\theta}$, $a = (0,\ldots,0,-1,1)^T$, $\hat{\theta} = (\hat{\beta}_0^T, \hat{\mu}_0, \hat{\mu}_1)^T$, $\hat{\mu}_0 = \Sigma_iT_iX_i\hat{\beta}_0/\Sigma_iT_i$, and $\hat{\mu}_1 = \Sigma_iT_iY_i/\Sigma_iT_i$. The estimating functions are $\psi_i = \psi_i = (S_{i0}^T, U_{i0}, U_{i1})^T$, where $(1-T_i)(Y_i - X_i^T\beta_0)X_i$, $U_{i0} = T_i(X_i\beta_0 - \mu_0)$, $U_{i1} = T_i(Y_i - \mu_1)$, and

$$A = \begin{bmatrix} A_{11} & 0 & 0 \\ A_{21} & A_{22} & 0 \\ 0 & 0 & A_{33} \end{bmatrix}.$$

The variance estimate is $N^{-1}a^T\hat{A}^{-1}\hat{B}\hat{A}^{-1T}a$, where $\hat{B} = N^{-1}\Sigma_i\hat{\psi}_i\hat{\psi}_i^T$, and the nonzero portions of $\hat{A}$ are $\hat{A}_{11} = N^{-1}\Sigma_i(1-T_i)X_iX_i^T$, $\hat{A}_{21} = -N^{-1}\Sigma_iT_iX_i^T$, and $\hat{A}_{22} = \hat{A}_{33} = N^{-1}\Sigma_iT_i$.

Similar arguments yield variance formulas for IPW estimates (Equations 20 and 21). Let $\gamma$ denote the unknown coefficients of the propensity model $\pi_i = (1 + \exp(-X_i^T\gamma))^{-1}$, and define $\mu_0 = E(Y_{i0})$, $\mu_1 = E(Y_{i1})$, and $\theta = (\gamma^T, \mu_0, \mu_1)^T$. We may write $\hat{ACE}$ *as* $\hat{\mu}_1 - \hat{\mu}_0 = a^T\hat{\theta}$, where $a = (0, \ldots, 0, -1, 1)^T$, $\hat{\theta} = (\hat{\gamma}^T, \hat{\mu}_0, \hat{\mu}_1)^T$, $\hat{\gamma}$ is the maximum-likelihood estimate for $\gamma$, $\hat{\mu}_0 = \Sigma_i(1-T_i)(1-\hat{\pi}_i)^{-1}Y_i/\Sigma_i(1-T_i)(1-\hat{\pi}_i)^{-1}$, and $\hat{\mu}_1 = \Sigma_iT_i\hat{\pi}_i^{-1}Y_i/\Sigma_iT_i\hat{\pi}_i^{-1}$ We may regard $\hat{\theta}$ as the solution to the joint estimating equations $\Sigma_i\psi_i(\theta) = (0\ldots,0)^T$, with $\psi_i(\theta) = \psi_i = (S_i^T, U_{i0}, U_{il})^T$, $S_i = (T_i - \pi_{iXi}$, $U_{i0} = (1-T_i)(1-\pi_i)^{-1}(Y_i - \mu_0)$, and $U_{i1} = T_i\pi_i^{-1}(Y_i - \mu_1)$. Note that $S_i$ is the score function for a logistic model (Ag, 2002). As $N \rightarrow \infty$, $\sqrt{N}(\hat{\theta} - \theta) \rightarrow N(0, \Sigma)$, where $\Sigma = A^{-1}BA^{-1T}$, $A = -E(\delta\psi_i/\delta\theta^T)$, and $B = E(\psi_i\psi_i^T)$. The matrix $A$ has the form

$$A = \begin{bmatrix} A_{11} & 0 & 0 \\ A_{21} & A_{22} & 0 \\ A_{31} & 0 & A_{33} \end{bmatrix},$$

where $A_{11} = -E(\delta S_i/\delta\gamma^T)$, $A_{21} = -E(\delta U_{i0}/\delta\gamma^T)$, $A_{22} = -E(\delta U_{i0}/\delta\mu_0)$, $A_{31} = -E(\delta U_{i1}/\delta\gamma^T)$, and $A_{33} = -E(\delta U_{i1}/\delta\mu_1)$. The variance of $\hat{ACE}$ can be approximated by $N^{-1}a^T\hat{A}^{-1}\hat{B}\hat{A}^{-1T}a$, where $\hat{B} = N^{-1}\Sigma_i\hat{\psi}_i\hat{\psi}_i^T$, $\hat{\psi}_i = \psi_i(\hat{\theta})$, and $\hat{A} = -N^{-1}\Sigma_i\delta\psi_i/\delta\theta^T$ evaluated at $\theta = \hat{\theta}$. The nonzero portions of $\hat{A}$ are

$$\hat{A}_{11} = N^{-1}\sum_i\hat{\pi}_i(1-\hat{\pi}_i)X_iX_i^T,$$

$$\hat{A}_{21} = -N^{-1}\sum_i(1-T_i)\hat{\pi}_i(1-\hat{\pi})^{-1}(Y_i - \hat{\mu}_0)X_i^T,$$

$$\hat{A}_{22} = N^{-1}\sum_i(1-T_i)(1-\hat{\pi}_i)^{-1},$$

$$\hat{A}_{31} = -N^{-1}\sum_iT_i\hat{\pi}_i^{-1}(1-\hat{\pi}_i)(Y_i - \hat{\mu}_1)X_i^T,$$

$$\hat{A}_{33} = -N^{-1}\sum_iT_i\hat{\pi}_i^{-1}.$$

Formulas for the IPW estimate of $ACE_1$ follow the same pattern, except now we define $\mu_0 = E(Y_{i0}|T_i = 1)$, $\mu_1 = E(Y_{i1}|T_i = 1)$,

(*Appendix continues*)

$$\hat{\mu}_0 = \frac{\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1}\hat{\pi}_i Y_i}{\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1}\hat{\pi}_1},$$

and $\hat{\mu}_1 = \sum_i T_i Y_i / \sum_i T_i$. Write $\hat{ACE}_1$ *as* $\hat{\mu}_1 - \hat{\mu}_0 = a^T\hat{\theta}$, where the estimating function for $\theta$ is $\psi_i = (S_i^T, U_{i0}, U_{i1})^T$ with $S_i = (T_i - \pi_i)X_i$, $U_{i0} = (1 - T_i)\pi_i(1 - \pi_i)^{-1}(Y_i - \mu_0)$, and $U_{i1} = T_i(Y_i - \mu_1)$. The estimated variance of $\hat{ACE}_1$ *is* $N^{-1}a^T\hat{A}^{-1}\hat{B}\hat{A}^{-1T}a$, where $\hat{B} = N^{-1}\sum_i \hat{\psi}_i\hat{\psi}_i^T$, and the relevant portions of $\hat{A}$ are $\hat{A}_{11} = N^{-1}\sum_i \hat{\pi}_i(1 - \hat{\pi}_i)X_iX_i^T$, $\hat{A}_{21} = -N^{-1}\sum_i (1 - T_i)\hat{\pi}_i(1 - \hat{\pi})^{-1}(Y_i - \hat{\mu}_0)X_i^T$, $\hat{A}_{22} = N^{-1}\sum_i (1 - T_i)\hat{\pi}_i(1 - \hat{\pi}_i)^{-1}$, $\hat{A}_{31}0$, and $\hat{A}_{33} = N^{-1}\sum_i T_i$.

For estimates based on dual models, our notation should allow different sets of covariates to enter the models for the propensities and the potential outcomes. Let us now use $Z_i$ to denote the predictors in the logistic model and $X_i$ the predictors in the linear models, so that $\pi_i = (1 + \exp(-Z_i^T\gamma))^{-1}$, and $E(Y_{it}|X_i) = X_i^T\beta_t$, $t = 0, 1$. The regression estimate for ACE with weighted residual bias corrections (Equation 22) can be written as $\hat{ACE} = (\hat{\mu}_1 + \hat{\xi}_1) - (\hat{\mu}_0 + \hat{\xi}_0)$, where $\hat{\mu}_1 = N^{-1}\sum_i X_i^T\hat{\beta}_1$, $\hat{\mu}_0 = N^{-1}\sum_i X_i^T\hat{\beta}_0$,

$$\hat{\xi}_1 = \frac{\sum_i T_i\hat{\pi}_i^{-1}(Y_i - X_i^T\hat{\beta}_1)}{\sum_i T_i\hat{\pi}_i^{-1}},$$

$$\hat{\xi}_0 = \frac{\sum_i (| - T_i\hat{\pi}_i)(| - \pi_i)^{-1}(Y_i - X_i^T\beta_0)}{\sum_i (| - T_i)(| - \pi_i)^{-1}}.$$

It is also $\hat{ACE} = a^T\hat{\theta}$, where $a = (0, \ldots, 0, -1, 1, -1, 1)^T$, and $\hat{\theta} = (\hat{\gamma}^T, \hat{\beta}_0^T, \hat{\beta}_1^T, \hat{\mu}_0, \hat{\mu}_1, \hat{\xi}_0, \hat{\xi}_1)^T$. The estimating functions for $\theta$ are $\psi_i = (W_i^T, S_{i0}^T, S_{i1}^T, U_{i0}, U_{i1}, V_{i0}, V_{i1})^T$, where $W_i = (T_i - \pi_i)Z_i$, $S_{i0} = (1 - T_i)(Y_i - X_i^T,\beta_0)X_{i1}$, $S_{i1} = T_i(Y_i - X_i^T\beta_1)X_i$, $U_{i0} = X_i^T\beta_0 - \mu_0$, $U_{i1} = X_i^T\beta_1 - \mu_1$, $V_{i0} = (1 - T_i)(1 - \pi_i)^{-1}(Y_i - X_i^T\beta_0 - \xi_0)$, and$V_{i0} = T_i\pi_i^{-1}(Y_i - X_i^T\beta_1 - \xi_1)$. The matrix $A = -E(\delta\psi_i/\delta\theta^T)$ can be partitioned as

$$A = \begin{bmatrix} A_{11} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_{22} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{33} & 0 & 0 & 0 & 0 \\ 0 & A_{42} & 0 & A_{44} & 0 & 0 & 0 \\ 0 & 0 & A_{53} & 0 & A_{55} & 0 & 0 \\ A_{61} & A_{62} & 0 & 0 & 0 & A_{66} & 0 \\ A_{71} & 0 & A_{73} & 0 & 0 & 0 & A_{77} \end{bmatrix},$$

and the variance estimate for $\hat{ACE}$ *is* $N^{-1}a^T\hat{A}^{-1}\hat{B}\hat{A}^{-1T}a$, where $\hat{B} = N^{-1}\sum_i \hat{\psi}_i\hat{\psi}_i^T$, and the nonzero portions of $\hat{A}$ are $\hat{A}_{11} = N^{-1}\sum_i \hat{\pi}_i(1 - \hat{\pi}_i)Z_iZ_i^T$, $\hat{A}_{22} = N^{-1}\sum_i (1 - T_i)X_iX_i^T$, $\hat{A}_{33} = N^{-1}\sum_i T_iX_iX_i^T$, $\hat{A}_{42} = \hat{A}_{53} = -N^{-1}\sum_i X_i^T$, $\hat{A}_{44} = \hat{A}_{55} = 1$,

$$\hat{A}_{61} = -N^{-1}\sum_i (1 - T_i)\hat{\pi}_i(1 - \hat{\pi}_i)^{-1}(Y_i - X_i^T\hat{\beta}_0 - \hat{\xi}_0)Z_i^T,$$

$$\hat{A}_{62} = N^{-1}\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1}X_i^T,$$

$$\hat{A}_{66} = N^{-1}\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1},$$

$$\hat{A}_{71} = N^{-1}\sum_i N^{-1}\sum_i T_i\hat{\pi}_i^{-1}(1 - \hat{\pi}_i)(Y_i - X_i^T\hat{\beta}_1 - \hat{\xi}_1)Z_i^T,$$

$\hat{A}_{73} = N^{-1}\sum_i T_i\hat{\pi}_i^{-1}X_i^T$ and $\hat{A}_{77} = N^{-1}\sum_i T_i\hat{\pi}_i^{-1}$.

Similarly, the weighted-residual estimate for $ACE_1$ (Equation 23) is $\hat{ACE}_1 = \hat{\mu}_1 - (\hat{\mu}_0 + \hat{\xi}_0)$, where $\hat{\mu}_1 = \sum_i T_iX_i^T\hat{\beta}_1/\sum_i T_i$, $\hat{\mu}_0 = \sum_i T_iX_i^T\hat{\beta}_0/\sum_i T_i$, and

$$\hat{\xi}_0 = \frac{\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1}\hat{\pi}_i(Y_i - X_i^T\hat{\beta}_0)}{\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1}\hat{\pi}_i}.$$

It can also be written as $\hat{ACE}_1 = a^T\hat{\theta}$, where $a = (0, \ldots, 0, -1, 1, -1)^T$, and $\hat{\theta} = (\hat{\gamma}^T, \hat{\beta}_0^T, \hat{\mu}_0, \hat{\mu}_1, \hat{\xi}_0)^T$. The estimating functions are $\psi_i = (W_i^T, S_{i0}^T, U_{i0}, U_{i1}, V_{i0})T$, $W_i = (T_i - \pi_i)Z_i$, $S_{i0} = (1 - T_i)(Y_i - X_i^T\beta_0)X_i$, $U_{i0} = T_i\,(X_{i\mathrm{T}\beta 0} - \mu_0)$, $U_{i1} = T_i(X_i^T\beta_1 - \mu_1)$, and $V_{i0} = (1 - T_i)(1 - \pi_i)^{-1}\pi_i(Y_i - X_i^T\beta_0 - \xi_0)$. The $A$ matrix is

$$A = \begin{bmatrix} A_{11} & 0 & 0 & 0 & 0 \\ 0 & A_{22} & 0 & 0 & 0 \\ 0 & A_{32} & A_{33} & 0 & 0 \\ 0 & 0 & 0 & A_{44} & 0 \\ A_{51} & A_{52} & 0 & 0 & A_{55} \end{bmatrix},$$

the variance estimate is $N^{-1}a^T\hat{A}^{-1}\hat{B}\hat{A}^{-1T}a$, $\hat{B} = N^{-1}\sum_i \hat{\psi}_i\hat{\psi}_i^T$, and the nonzero portions of $\hat{A}$ are $\hat{A}_{11} = N^{-1}\sum_i \hat{\pi}_i(1 - \hat{\pi}_i)Z_iZ_i^T$, $\hat{A}_{22} = N^{-1}\sum_i (1 - T_i)X_iX_i^T$, $\hat{A}_{32} = -N^{-1}\sum_i T_iX_i^T$, $\hat{A}_{33} = \hat{A}_{44} = N^{-1}\sum_i T_i$,

$$\hat{A}_{51} = -N^{-1}\sum_i (1 - T_i)\hat{\pi}_i(1 - \hat{\pi}_i)^{-1}(Y_i - X_i^T\hat{\beta}_0 - \hat{\xi}_0)Z_i^T,$$

$$\hat{A}_{52} = N^{-1}\sum_i (1 - T_i)\hat{\pi}_i(1 - \hat{\pi}_i)^{-1}X_i^T,$$

and $\hat{A}_{55} = N^{-1}\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1}\hat{\pi}_i$.

For the weighted regression estimate of ACE, take $\hat{\mu}_0 = N^{-1}\sum_i X_i^T\hat{\beta}_0$, and $\hat{\mu}_1 = N^{-1}\sum_i X_i^T\hat{\beta}_1$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the weighted least-squares estimates from Equations 25 and 24. Then $\hat{ACE} = a^T\hat{\theta}$, where $a = (0, \ldots, 0, -1, 1)^T$, and $\hat{\theta} = (\hat{\gamma}^T, \hat{\beta}_0^T, \hat{\beta}_1^T, \hat{\mu}_0^{\mu_1})^T$. The estimating functions are $\psi_i = (W_i^T, S_{i0}^T, S_{i1}^T, U_{i0}, U_{\mathrm{lsi}1})^T$, $W_i = (T_i - \pi_i)Z_i$, $S_{i0} = (1 - T_i)(1 - \pi_i)^{-1}(Y_i - X_i^T\beta_0)X_i$, $S_{i1} = T_i\pi^{-1}{}_i(Y_i - X_i^T\beta_1)X_i$, $U_{i0} = X_i^T\beta_0 - \mu_0$, and $U_{i1} = X_i^T\beta_1 - \mu_1$. The pattern of $A$ is

$$A = \begin{bmatrix} A_{11} & 0 & 0 & 0 & 0 \\ A_{21} & A_{22} & 0 & 0 & 0 \\ A_{31} & 0 & A_{33} & 0 & 0 \\ 0 & A_{42} & 0 & A_{44} & 0 \\ 0 & 0 & A_{53} & 0 & A_{55} \end{bmatrix},$$

the variance estimate is $N^{-1}a^T\hat{A}^{-1}\hat{B}\hat{A}^{-1T}a$, $\hat{B} = N^{-1}\sum_i\hat{\psi}_i\hat{\psi}_i^T$, and the parts of $\hat{A}$ are $\hat{A}_{11} = N^{-1}\sum_i\hat{\pi}_i(1 - \hat{\pi}_i)Z_iZ_i^T$, $\hat{A}_{21} = -N^{-1}\sum_i(1 - T_i)\hat{\pi}_i(1 - \hat{\pi}_i)^{-1}(Y_i - X_i^T\hat{\beta}_0)X_iZ_i^T$, $\hat{A}_{22}$ $N^{-1}\sum_i(1 - T_i)(1 - \hat{\pi}_i)^{-1}X_iX_i^T$, $\hat{A}_{31} = N^{-1}\sum_iT_i\hat{\pi}_i^{-1}(1 - \hat{\pi}_i)(Y_i - X_i^T\hat{\beta}_1)X_iZ_i^T$, $\hat{A}_{33}$ $N^{-1}\sum_iT_i\hat{\pi}_i^{-1}X_iX_i^T$, $\hat{A}_{42} = \hat{A}_{53} = -N^{-1}\sum_iX_i^T$, and $\hat{A}_{44} = \hat{A}_{55} = 1$.

Finally, the weighted regression estimate $A\hat{C}E_1$ *is* $\hat{\mu}_1 - \hat{\mu}_0$, where $\hat{\mu}_1 = \sum_iT_iY_i/\sum_iT_i$, $\hat{\mu}_0 = \sum_iT_iX_i^T\hat{\beta}_0/\sum_iT_i$, and

$$\hat{\beta}_0 = \left(\sum_i(1 - T_i)(1 - \hat{\pi})^{-1}\hat{\pi}_iX_iX_i^T\right)^{-1} \times \left(\sum_i(1 - T_i)(1 - \hat{\pi}_i)^{-1}\hat{\pi}_iX_iY_i\right).$$

Take $A\hat{C}E = a^T\hat{\theta}$, $a = (0,\ldots,0, -1, 1)^T$, $\hat{\theta} = (\hat{\gamma}^T, \hat{\beta}_0{}^T, \hat{\mu}_0, \hat{\mu}_1)^T$. The estimating functions are $\psi_i = (W_i^T, S_{i0}{}^T, U_{i0}, U_{i1})^T$, $W_i = (T_i - \pi_i)Z_i$, $S_{i0} = (1 - T_i)(1 - \pi_i)^{-1}\pi_i(Y_i - X_i{}^T\beta_0)X_i$, $U_{i0} = T_i(X_i{}^T\beta_0 - \mu_0)$, and $U_{i1} = T_i(Y_i - \mu_1)$. The $A$ matrix is

$$A = \begin{bmatrix} A_{11} & 0 & 0 & 0 \\ A_{21} & A_{22} & 0 & 0 \\ 0 & A_{32} & A_{33} & 0 \\ 0 & 0 & 0 & A_{44} \end{bmatrix},$$

the variance estimate is $N^{-1}a^T\hat{A}^{-1}\hat{B}\hat{A}^{-1T}a$, $\hat{\beta} = N^{-1}\sum_i\hat{\psi}_i\hat{\psi}_i^T$, and the relevant parts of $\hat{A}$ are $\hat{A}_{11} = N^{-1}\sum_i\hat{\pi}_i(1 - \hat{\pi}_i)Z_iZ_i^T$, $\hat{A}_{21} = -N^{-1}\sum_i(1 - T_i)\hat{\pi}_i(1 - \hat{\pi}_i)^{-1}(Y_i - X_i^T\hat{\beta}_0)X_iZ_i^T$, $\hat{A}_{22} = N^{-1}\sum_i(1 - T_i)(1 - \hat{\pi}_i)^{-1}\hat{\pi}_iX_iX_i^T$, $\hat{A}_{32} = N^{-1}\sum_iT_iX_i^T$, and $\hat{A}_{33} = \hat{A}_{44} = N^{-1}\sum_iT_i$.

Received July 21, 2008
Revision received September 26, 2008
Accepted September 30, 2008 ■