

**Matching Estimators of Causal Effects:
From Stratification and Weighting
to Practical Data Analysis Routines ***

Stephen L. Morgan
Department of Sociology
358 Uris Hall
Cornell University
Ithaca, NY 14850
(slm45@cornell.edu)

David J. Harding
Department of Sociology
William James Hall, 5th Floor
Harvard University
Cambridge, MA 02138
(dharding@wjh.harvard.edu)

Key words: causality, counterfactuals, observational study, propensity score, treatment effect

Shortened title: Matching Estimators

Corresponding author: Stephen Morgan. Address and email above are correct. Phone number is (607) 255-0706, and FAX is (607) 255-8473.

* We thank Sascha Becker, Edwin Leuven, Herb Smith, Elizabeth Stuart, and Yu Xie for helpful comments.

**Matching Estimators of Causal Effects:
From Stratification and Weighting
to Practical Data Analysis Routines**

ABSTRACT

As the counterfactual model of causality has increased in popularity, sociologists have returned to matching as a research methodology. In this paper, advances over the past two decades in matching estimators are explained. After presenting the fundamental concepts of the counterfactual model of causality, we introduce matching methods by focusing first on ideal scenarios in which stratification and reweighting procedures can warrant causal inference. Then, we discuss how matching is often undertaken in practice, offering an overview of various algorithms. Finally, we discuss how the assumptions behind matching estimators can break down in practice. In conclusion, we outline some practical advice on matching, and discuss the combination of matching with regression methods.

INTRODUCTION

The counterfactual, or “potential outcomes,” model of causality offers new possibilities for the formulation and investigation of causal questions in sociology. In the language of Holland (1986), the counterfactual perspective shifts attention from the identification of the “causes of effects” toward the more tractable goal of estimating the “effects of causes.” Accordingly, the primary goal of causal analysis becomes the investigation of selected effects of a particular cause, rather than the search for all possible causes of a particular outcome along with the comprehensive estimation of all of their relative effects.

The rise of the counterfactual model to prominence has increased the popularity of data analysis routines that are most clearly useful for estimating the effects of causes. The matching estimators that we will review and explain in this article are perhaps the best example of a classic research design that has been re-born in the past two decades as a data analysis procedure for estimating causal effects. Matching represents an intuitive method for addressing causal questions, primarily because it pushes the analyst to confront the process of causal exposure as well as the limitations of available data. Accordingly, among social scientists who adopt a counterfactual perspective, matching methods (particularly propensity score matching) are fast becoming an indispensable technique for prosecuting causal questions, even though they usually prove to be the beginning rather than the end of causal analysis on any particular topic.

Yet, while empirical examples that demonstrate the utility of matching methods have begun to accumulate, the methodological literature has fallen behind in providing an up-to-date treatment of both the fundamentals of matching as well as the recent developments in practical matching methodology. Thus, the purpose of this article is to provide a starting point for those sociologists who are sophisticated users of other quantitative methods and who want to

understand matching methods, either to become savvy consumers of existing research using matching methods, to begin using matching methods in their own work, or to teach matching methods in graduate methods courses. Although our agenda is primarily explanatory, we will also make the case that matching techniques should be used as causal effect estimators in sociology more frequently than is now the case, but still with appropriate caution and not to the exclusion of other more established methods.

We begin with a brief discussion of the past use of matching methods. Some sociologists may be surprised to learn that the sociological literature contains many of the early developments of matching methods. We then outline the key ideas of the counterfactual model of causality, with which most matching methods are now motivated. Then, we present the fundamental concepts underlying matching, including stratification of the data, overlapping supports, reweighting, and propensity scores. Thereafter, we discuss how matching is undertaken in practice, including an overview of various matching algorithms. Finally, we discuss how the assumptions behind matching estimators often break down in practice, along with some of the remedies that have been proposed to address the resulting problems. In the course of presentation, we offer four hypothetical examples which demonstrate some essential results from the matching literature, progressing from idealized examples of stratification and weighting to the implementation of alternative matching algorithms on simulated data where the treatment effects of interest are known by construction.

ORIGINS AND MOTIVATIONS FOR MATCHING ESTIMATORS

Matching techniques have origins in experimental work from the first half of the twentieth

century. Relatively sophisticated discussions of matching as a research design can be found in early methodological texts in sociology (see Greenwood 1945) and also in attempts to adjudicate between competing explanatory accounts in applied demography (Freedman and Hawley 1949). This early work continued in sociology (e.g., Althausser and Rubin 1970, 1971; Yinger et al. 1967) right up to the key foundational literature (Rubin 1973a, 1973b, 1976a, 1976b, 1977, 1979, 1980) for the new wave of matching techniques that we will present in this article.

In the early 1980s, matching techniques, as we conceive of them now, were advanced in a set of papers by Rosenbaum and Rubin (1983a, 1984, 1985a, b) that offered solutions to a variety of practical problems which until then limited matching techniques to very simple applications. Variants of these new techniques found some use immediately in sociology (Berk and Newton 1985; Berk et al. 1986; Hoffer et al. 1985), continuing with work by Smith (1997). In the late 1990s, economists joined in the development of matching techniques in the course of evaluating social programs (e.g., Heckman et al. 1997, 1998a, b, 1999). New sociological applications are now accumulating (DiPrete and Engelhardt 2004; DiPrete and Gangl 2004; Harding 2003; Morgan 2001), and we suspect that matching will complement other types of modeling in sociology with much greater frequency in the near future.

In this literature, matching is usually introduced in one of two ways: (1) as a method to form quasi-experimental contrasts by sampling comparable treatment and control cases from among a larger pool of such cases or (2) as a non-parametric method of adjustment for treatment assignment patterns when it is feared that ostensibly simple parametric regression estimators cannot be trusted.

For the first motivation, the archetypical example is an observational biomedical study

where a researcher is called upon to assess what can be learned about a particular treatment which has been given haphazardly (but which, according to the testimony of health professionals, shows some signs of effectiveness). The investigator is given access to medical records with: (1) a measurement of initial symptoms, $Y_{i,t=0}$, (2) an indicator variable for whether the treatment, D_i , was taken, (3) a subsequent measurement of symptoms from a return visit, $Y_{i,t=1}$, and (4) a set of characteristics of individuals, as a vector of variables \mathbf{X}_i , that are drawn from demographic profiles and health histories prior to the onset of initial symptoms. Typically, the cases are not drawn from a population via any known sampling scheme, as the data emerge first as a result of the distribution of initial symptoms and then as a function of access to the relevant health clinic. And, moreover, the treatment is scarce in most examples, such that only a small proportion of those with some symptoms were given the treatment.

For examples such as this one, matching is used to select a non-treated case for each treated case based on identical initial symptoms, $Y_{i,t=0}$, and other characteristics, \mathbf{X}_i . All treated cases and matched cases are retained, but all non-matched cases are discarded. Differences in $Y_{i,t=1}$ are then calculated for treated and matched cases, with the average difference serving as the treatment effect estimate for the group of individuals given the treatment.¹

The second form of motivation has no archetypical example, as it is nearly identical to the generalized motivation for using regression analysis as a technique for estimating a binary causal effect. Suppose, for example, that an investigator is interested in the causal effect of an observed

¹ A virtue of matching, as developed in this tradition, is cost-effectiveness for prospective studies. If the goal of a study is to measure the evolution of a causal effect over time by measuring symptoms at several points in the future, then discarding non-treated cases unlike any treated cases can cut expenses without substantially affecting the quality of causal inferences that a study can yield.

dummy variable, D_i , on an observed outcome Y_i . Further suppose that it is known that a simple bivariate regression, $Y_i = \alpha + \gamma D_i + \varepsilon_i$, will yield an estimated coefficient $\hat{\gamma}$ that is a biased estimate of the true causal effect because the causal variable D_i is associated with omitted variables embedded in the error term, ε_i . For example, if D_i is the receipt of a college degree and Y_i is a measure of economic success, then the goal of the research is to determine the effect of obtaining a college degree on subsequent economic success. However, family background variables are correlated with both D_i and Y_i , introducing classical omitted variables bias into the estimation of the bivariate regression coefficient for college degrees.

In general, for this second motivation of matching, a set of variables in \mathbf{X}_i (i.e., family background for the college degree example) is assumed to predict both D_i and Y_i , but the non-equivalence in \mathbf{X}_i across levels of the cause D_i creates complications for the analyst. One solution is to estimate an expanded regression model by including these variables in an equation: $Y_i = \alpha + \gamma D_i + \boldsymbol{\beta}'\mathbf{X}_i + \varepsilon_i$. With this strategy, the goal is to simultaneously estimate the causal effects of both \mathbf{X}_i and D_i on the outcome, Y_i .

In this context, a matching estimator is nothing other than an alternative to parametric regression adjustment. Rather than attempt to estimate the causal effects of both \mathbf{X}_i and D_i with a parametric model, a matching estimator non-parametrically “balances” the variables in \mathbf{X}_i across levels of D_i solely in the service of obtaining the best possible estimate of the causal effect of D_i . As we will describe later, the most popular technique is to estimate the probability of D_i as a function of \mathbf{X}_i (usually using a logit or probit model), and then match cases based on the predicted probability from this model. This procedure results in a subsampling of cases across levels of D_i , comparable to the matching procedure described for the biomedical example, but for

a single dimension that is a function of the variables in \mathbf{X}_i .

One virtue of the methodological literature on matching is the amount of attention that has been given to specific examples in the applied literature. This attention has, however, come at a cost, as the weaknesses of the techniques are sometimes downplayed so that the potential of the techniques can be promoted. In the remainder of this article, we will invert the usual presentation of matching, starting first with simple idealized scenarios in which matching via stratification and reweighting solves all problems of causal inference. Then, we will progressively add real-world complexity to the presentation, so as to explain how such complexity can undermine the power of the techniques.

COUNTERFACTUALS AND CAUSAL EFFECTS

Although matching can be seen as an extension of tabular analysis of simple three-way cross-classifications (i.e., causal variable by outcome variable by adjustment variable), the current literature is primarily associated with counterfactual models of causality.² Accordingly, from here onward, we will adopt the language that dominates this framework, writing of the causal variable of interest as a treatment variable. And, as will become apparent later, we confine most of our attention to binary treatments, generally referring to the group that receives the treatment as the

² See Winship and Morgan (1999) and Sobel (1995) for presentations of the counterfactual model in sociology. In this article, we adopt the foundational assumptions of the literature on counterfactual causality, such as the stable unit treatment value assumption, which stipulates that the causal effect for each individual does not depend on the treatment status of any other individual in the population. When this non-independence assumption is violated, complications beyond the scope of this article arise.

treatment group and the group that does not as the control group.³ One could, however, re-write all that follows referring to such groups as those who are exposed to the cause and those who are not.

In the counterfactual framework, we approach causal inference by first stipulating the existence of two potential outcome random variables that are defined over all individuals in the population. Y^t is the potential outcome in the treatment state, and Y^c is the potential outcome in the control state. The individual-level causal effect of the treatment is then defined as:

$$(1) \quad \delta_i = Y_i^t - Y_i^c.$$

Because we can never observe the potential outcome under the treatment state for those observed in the control state (and vice versa), we can never know the individual-level causal effect in Equation 1.⁴ This predicament is sometimes labeled the fundamental problem of causal inference (Holland 1986).

We can only observe the variable Y_i , which is related to the potential outcomes by:

³ For extensions of matching to multi-valued causal/treatment variables, see Angrist and Krueger (1999), Hirano and Imbens (2004), Imbens (2000), Lechner (2002a, b), Lu et al. (2001), Rosenbaum (2002), and Imai and van Dyk (2004). As one will see from reading this literature, the added complexity presented by multi-valued and continuous treatments can be considerable, to the extent that matching loses much of its transparency and is then no more intuitive than regression (and, because of its unfamiliarity, then appears vastly more complex than regression). For these reasons, for the foreseeable future we expect that most applied researchers will use matching only for the estimation of binary causal effects. Since such effects are usually the primitives of all more encompassing multi-valued treatment effects, this may not be as severe of a limitation as one might fear.

⁴ There is a wide variety of notation in the potential outcome literature, and we have adopted notation that we feel is the easiest to grasp. Equation 1 is often written as one of the following alternatives: $\delta_i = Y_{it} - Y_{ic}$, $\Delta_i = Y_{it} - Y_{0i}$, $\tau_i = Y_i(1) - Y_i(0)$, or $f_i(1) - f_i(0)$. We therefore use the right superscript to denote the potential treatment state of the corresponding potential outcome variable. To make this more explicit, we could define a general treatment variable (or causal variable), X , and then write Equation 1 as: $\delta_i = Y_i^{X=t} - Y_i^{X=c}$.

$$Y_i = Y_i^t \quad \text{if} \quad T_i = 1$$

$$Y_i = Y_i^c \quad \text{if} \quad T_i = 0 ,$$

where the binary variable, T_i , is equal to 1 if an individual receives the treatment (i.e., is exposed to the cause) and equal to 0 if an individual receives the control (i.e., is not exposed to the cause).

This paired definition is generally written compactly as:

$$(2) \quad Y_i = T_i Y_i^t + (1 - T_i) Y_i^c .$$

Because it is usually impossible to estimate individual-level causal effects, we typically shift attention to aggregated causal effects, such as the average causal effect:

$$(3) \quad E[\delta_i] = E[Y_i^t] - E[Y_i^c] ,$$

where $E[\cdot]$ is the expectation operator from probability theory, denoting the expected value of the random variable or expression that is its argument.

The naive estimator of the average causal effect is:

$$(4) \quad E_N[Y_i | T_i = 1] - E_N[Y_i | T_i = 0] ,$$

where $E_N[\cdot]$ denotes the sample expectation of its argument for a sample of size N . In other words, the subscript N serves the same basic notational function as an overbar on Y_i . We use this notation, as it allows for greater clarity in aligning sample and population-level conditional expectations for subsequent expressions.

When does the naive estimator yield an unbiased and consistent estimate of the average treatment effect? First, decompose the average causal effect as:

$$(5) \quad E[\delta_i] = \{ \pi E[Y_i^t | T_i = 1] + (1 - \pi) E[Y_i^t | T_i = 0] \} \\ - \{ \pi E[Y_i^c | T_i = 1] + (1 - \pi) E[Y_i^c | T_i = 0] \} .$$

where π is the proportion of the population that would take the treatment instead of the control (under whatever treatment selection regime or causal exposure mechanism prevails in the particular application). The average treatment effect is then a function of five unknowns: the proportion of the population that would be assigned to the treatment along with four conditional expectations of the potential outcomes. Without introducing additional assumptions about the treatment assignment/selection mechanism, we can effectively estimate with observational data from a sample of the population only three of the five unknowns on the right-hand side of Equation 5.

We know that, for a very large random sample, the mean of the dummy treatment variable T_i would be equal to the true proportion of the population that would be assigned to (or would select into) the treatment. More precisely, we know that the sample mean of T_i converges in probability to π , which we write as:

$$(A1) \quad E_N[T_i] \xrightarrow{p} \pi .$$

Although the notation of Assumption A1 may appear unfamiliar, the claim is that, as the sample size, N , increases, the sample mean of T_i approaches the true value π , which is a fixed population parameter. We can offer similar claims about two other unknowns:

$$(A2) \quad E_N[Y_i | T_i = 1] \xrightarrow{p} E[Y_i^t | T_i = 1] \quad \text{and}$$

$$(A3) \quad E_N[Y_i | T_i = 0] \xrightarrow{p} E[Y_i^c | T_i = 0] ,$$

indicating that the sample mean of the observed outcome in the treatment group converges to the true average outcome under the treatment state for those in the treatment group (and analogously

for the control group and control state).⁵

Unfortunately, however, there is no generally effective way to estimate the two remaining unknowns in Equation 5: $E[Y_i^t | T_i = 0]$ and $E[Y_i^c | T_i = 1]$. These are counterfactual expectations: the average outcome under the treatment for those in the control group and the average outcome under the control for those in the treatment group. Without further assumptions, no estimated quantity based on observed data would converge to the true values for these unknowns.

What assumptions would suffice to enable consistent estimation of the average treatment effect with observed data? If we could randomize treatment assignment, we could introduce the assumptions:

$$(A4) \quad E[Y_i^t | T_i = 1] = E[Y_i^t | T_i = 0] \text{ and}$$

$$(A5) \quad E[Y_i^c | T_i = 1] = E[Y_i^c | T_i = 0]$$

and then substitute into Equation 5 in order to reduce the number of unknowns from the original five parameters to the three parameters that we know from Assumptions A1 through A3 can be consistently estimated with the data. When randomization is infeasible, as is the case with most applications in the social sciences, such simplification through substitution is impossible.

Although the unconditional average treatment effect is the most common subject of investigation in sociology, more narrowly defined average treatments are of interest as well, as we show in the examples later. The average treatment effect for those who take the treatment is:

$$(6) \quad E[\delta_i | T_i = 1] = E[Y_i^t | T_i = 1] - E[Y_i^c | T_i = 1],$$

⁵ Assumptions A1 through A3 are fairly obvious, and thus the convergence notation may well be unnecessary. We err on the side of precision of notation at this point because, as we will show later, much of the confusion over the power of matching arises from a lack of appreciation for the different problems created by sparseness of data and sampling error relative to more serious forms of data insufficiency.

and the average treatment effect for those who do not take the treatment is:

$$(7) \quad E[\delta_i | T_i = 0] = E[Y_i^t | T_i = 0] - E[Y_i^c | T_i = 0].$$

As will become clear, in many cases only one of the two average treatments effects in Equations 6 and 7 can be estimated consistently, and when this is the case the overall average treatment effect in Equation 3 cannot be estimated consistently. Other average causal effects (or more general properties of the distribution of causal effects) are often of interest as well, and Heckman (2000), Manski (1995), Rosenbaum (2002), and Pearl (2000) all give full discussions of the variety of causal effects which may be relevant for different types of applications. In this article, we will focus almost exclusively on the three types of average causal effects represented by Equations 3, 6, and 7.

STRATIFICATION AND PROPENSITY SCORES

In this section, we introduce matching estimators in idealized research conditions, focusing first on stratification of the data and then explaining the theoretical utility of propensity scores.

Thereafter, we proceed to a discussion of matching in more realistic scenarios, which is where we explain the developments of matching techniques which have been achieved in the last decade.

Estimating Causal Effects by Stratification

Suppose that those who take the treatment and those who do not are very much unlike each other, and yet the ways in which they differ are captured exhaustively by a set of observed treatment assignment/selection variables collectively stored in a vector \mathbf{S}_i . For the language we will adopt in this article, knowledge and observation of \mathbf{S}_i allows for a “perfect stratification” of

the data. By “perfect,” we mean precisely that individuals within groups defined by values on the variables in \mathbf{S}_i are entirely indistinguishable from each other in all ways except for (1) observed treatment status and (2) completely random shocks to the potential outcome variables. With such a perfect stratification of the data, even though we would not be able to assert Assumptions A4 and A5, we would be able to assert conditional variants of those assumptions:

$$(A4-S) \quad E[Y_i^t | T_i = 1, \mathbf{S}_i] = E[Y_i^t | T_i = 0, \mathbf{S}_i].$$

$$(A5-S) \quad E[Y_i^c | T_i = 1, \mathbf{S}_i] = E[Y_i^c | T_i = 0, \mathbf{S}_i].$$

These assumptions would suffice to enable consistent estimation of the average treatment effect, as the treatment can be considered randomly assigned within groups defined by values on the variables in \mathbf{S}_i .⁶

Before we introduce an idealized example of perfect stratification, first note why everything works out so cleanly when a set of stratifying variables is available. If A4-S is valid,

$$\begin{aligned} \text{then (8)} \quad E[\delta_i | T_i = 0, \mathbf{S}_i] &= E[Y_i^t | T_i = 0, \mathbf{S}_i] - E[Y_i^c | T_i = 0, \mathbf{S}_i] \\ &= E[Y_i^t | T_i = 1, \mathbf{S}_i] - E[Y_i^c | T_i = 0, \mathbf{S}_i]. \end{aligned}$$

If A5-S is valid, then

$$\begin{aligned} \text{(9)} \quad E[\delta_i | T_i = 1, \mathbf{S}_i] &= E[Y_i^t | T_i = 1, \mathbf{S}_i] - E[Y_i^c | T_i = 1, \mathbf{S}_i] \\ &= E[Y_i^t | T_i = 1, \mathbf{S}_i] - E[Y_i^c | T_i = 0, \mathbf{S}_i]. \end{aligned}$$

⁶ When in this situation, researchers often argue that the naive estimator is subject to bias (either generic omitted variable bias or individually generated selection bias). But, since a perfect stratification of the data can be formulated, the study is said to be free of hidden bias (see Rosenbaum 2002), treatment assignment is ignorable (see Rosenbaum and Rubin 1983a), or treatment selection is on the observable variables only (Heckman et al. 1999). Rosenbaum (2002) stresses the utility of asserting a no-hidden-bias assumption in an observational study but not then succumbing to over-confidence. The assumption allows one to obtain a causal effect estimate, but the initial estimate must be interpreted with caution and examined for its sensitivity to reasonable violations of assumptions A4-S and A5-S.

Both of the last two lines of Equations 8 and 9 are identical, and neither includes counterfactual conditional expectations. One can therefore estimate effectively Equations 8 and 9 if these assumptions hold, and thus obtain consistent estimates of treatment effects conditional on \mathbf{S}_i . To then form consistent estimates of alternative treatment effects, one simply averages the stratified estimates over the distribution of \mathbf{S}_i , as we show in the following hypothetical example.

Hypothetical Example 1

Consider an example where we have data that yield $E_N[Y_i | T_i = 1] = 10.2$ and $E_N[Y_i | T_i = 0] = 4.4$. The naive estimate of the average causal effect is therefore 5.8, which is the difference between the average outcome observed for the treatment group and the average outcome observed for the control group (i.e., $10.2 - 4.4$). Suppose that we suspect, based on theory, that this is a poor estimate of the causal effect of the treatment. But, suppose that we have a single perfectly stratifying variable (as in Rubin 1977) that has three categories. Moreover, suppose, for simplicity of exposition, that our sample is large enough such that sampling error is trivial. Therefore, we can assume that the sample moments in our data match the population moments (i.e., $E[Y_i^t | T_i = 1] = 10.2$ and $E[Y_i^c | T_i = 0] = 4.4$, etc.).

Consider, now, an underlying set of potential outcome variables and treatment assignment patterns that could give rise to a naive estimate of 10.2. Table 1 gives the joint probability distribution of the treatment variable T and the stratifying variable S in its first panel as well as expectations, conditional on S , of the potential outcomes under the treatment and control states. The joint distribution in the first panel shows that individuals with S equal to 1 are more likely to be observed in the control group, individuals with S equal to 2 are

equally likely to be observed in the control group and the treatment group, and individuals with S equal to 3 are more likely to be observed in the treatment group.

[INSERT TABLE 1 ABOUT HERE]

As shown in the second panel of Table 1, the average potential outcomes conditional on S and T imply that the average causal effect is 2 for those with S equal to 1 or S equal to 2 but 4 for those with S equal to 3 (see the last column). Moreover, as shown in the last row of the table, where the potential outcomes are averaged over the within- T distribution of S , $E[Y_i^t | T_i = 1] = 10.2$ and $E[Y_i^c | T_i = 0] = 4.4$, matching the initial set-up of the example based on a naive estimate of 5.8 from a very large sample.

Table 2 shows what can be calculated from the data, assuming that S offers a perfect stratification of the data. The first panel presents the sample expectations of the observed outcome variable conditional on T and S . The second panel of Table 2 presents corresponding sample estimates of the conditional probabilities of S given T . The estimated values are for a very large sample, as stipulated earlier, such that sampling error is infinitesimal.

[INSERT TABLE 2 ABOUT HERE]

The existence of a perfect stratification assures that the conditional expectations in the first panel of Table 2 match those of the second panel of Table 1. When stratifying the population by S , the average observed outcome for those in the control/treatment group with a particular value of S is equal to the average potential outcome under the control/treatment state for those with a particular value of S . Conversely, if S were not a perfect stratifying variable, then the sample means in the first panel of Table 2 would not converge to the population expectations of the potential outcomes in the second panel of Table 1. The sample

means would be based on heterogeneous groups of individuals who differ systematically within the strata defined by S in ways that are correlated with individual-level treatment effects.

If S offers a perfect stratification of the data, then, with a suitably large sample, one can estimate from the numbers in the cells of the two panels of Table 2 both the average treatment effect among the treated as $(4-2)(.2) + (8-6)(.3) + (14-10)(.5) = 3$ and the average treatment effect among the untreated as $(4-2)(.6) + (8-6)(.2) + (14-10)(.2) = 2.4$. Finally, if one were to calculate the appropriate marginal distributions of S and T (using sample analogs for the marginal distribution from the first panel of Table 1), one can perfectly estimate the unconditional average treatment effect either as $(4-2)(.44) + (8-6)(.24) + (14-10)(.32) = 2.64$ or as $3(.6) + 2.4(.4) = 2.64$. Thus, for this hypothetical example, the naive estimator would be asymptotically upwardly biased for the average treatment effect among the treated, the average treatment effect among the untreated, and the unconditional average treatment effect. But, by appropriately weighting the stratified estimates of the treatment effect, consistent estimates of the average treatment effects in Equations 3, 6, and 7 can be obtained.

In general, if a stratifying variable S completely accounts for all systematic differences between those who take the treatment and those who do not, then conditional-on- S estimators yield consistent estimates of the average treatment effect conditional on S :

$$\{E_N[Y_i|T_i = 1, S_i = s] - E_N[Y_i|T_i = 0, S_i = s]\} \xrightarrow{p} E[Y_i^t - Y_i^c | S_i = s] = E[\delta_i | S_i = s] .$$

One can then take weighted sums of these stratified estimators, such as for the unconditional average treatment effect:

$$\sum_s \{E_N[Y_i|T_i = 1, S_i = s] - E_N[Y_i|T_i = 0, S_i = s]\} \Pr_N[S_i = s] \xrightarrow{p} E[\delta_i] .$$

Substituting into this last expression the distributions of S conditional on the two possible values of T , one can obtain consistent estimates of the average treatment effect among the treated and the average treatment effect among the untreated.

The key to using stratification to solve the causal inference problem for all three causal effects of primary interest is twofold: finding the stratifying variable and then obtaining the marginal probability distribution $\Pr(S)$ as well as the conditional probability distribution $\Pr(S|T)$. Once these two steps are accomplished, obtaining consistent estimates of the within-strata treatment effects is straightforward, and one then simply weights the stratified estimates accordingly. Indeed, if a perfect stratification of the data can be found, the data can be analyzed as if they are a stratified random sample with the treatment randomly assigned within each stratum. Even the variance estimates from stratified sampling carry over (see Rosenbaum 2002, Chap. 3).

Support Conditions for Stratifying Variables

Suppose now that a perfect stratification of the data is available, but that there is a stratum in which no member of the population ever receives the treatment. Here, the average treatment effect is undefined. A hidden stipulation is built into Assumptions A4-S and A5-S if one wishes to be able to estimate the average treatment effect for the entire population. The “perfect” stratifying variables must not be so perfect that they sort deterministically all individuals into the treatment and the control groups. If so, the support of the stratifying variables differs for treatment and control cases, necessitating a re-definition of the causal effect of interest.

Hypothetical Example 2

For the example depicted in Tables 3 and 4, S again offers a perfect stratification of the data. The set-up of these two tables is exactly equivalent to the prior Tables 1 and 2, respectively. The major difference is evident in the joint distribution of S and T presented in the first panel of Table 3. As shown in the first cell of the second column, no individual with S equal to 1 would ever be observed in the treatment group of a data set of any size, as the joint probability of S equal to 1 and T equal to 1 is zero. Corresponding to this structural zero in the joint distribution of S and T , the second panel of Table 3 shows that there is no corresponding potential outcome under the treatment statement for those with S equal to 1. And, thus, as shown in the last column of the second panel of Table 3, no causal effect exists for individuals with S equal to 1.⁷

[INSERT TABLES 3 AND 4 ABOUT HERE]

Table 4 shows what can be estimated from a very large sample for this example. If S offers a perfect stratification of the data, one could consistently estimate the treatment effect for the treated as $(8-6)(.325) + (14-10)(.675) = 3.35$. There is, unfortunately, no way to consistently estimate the treatment effect for the untreated, and hence no way to consistently estimate the unconditional average treatment effect.

Are examples such as this one ever found in practice? Consider the evaluation of a generic program in which there is an eligibility rule. One simply cannot estimate the likely benefits

⁷ The naive estimator can be calculated for this example, and it would equal 8.05 for a very large sample because $[8(.325) + 14(.675)] - [2(.667) + 6(.167) + 10(.167)]$ is equal to 8.05. See the last row of Table 3 for the population analogs to the two pieces of the naive estimator.

of enrolling in the program for those who are ineligible, even though, if some of those individuals were enrolled in the program they would likely be affected by the treatment in some way (but, of course, in a way that may be very different from those who do enroll in the program).

Perhaps the most important point of this last example, however, is that the naive estimator is entirely misguided for this hypothetical application. The average treatment effect is undefined. More generally, not all causal questions have answers worth seeking even in best-case data availability scenarios, and sometimes this will be clear from the data and contextual knowledge of the application. However, at other times, the data may appear to suggest that no causal inference is possible even though the problem is simply a small sample size. There is a clever solution to sparseness of data for these types of situations, which we discuss in the next section.

Estimating Causal Effects Using Propensity Scores

Most matching estimators rely on what have become known as estimated propensity scores – the estimated probability of taking the treatment as a function of variables that predict treatment assignment. Before explaining the attraction of estimated propensity scores, there is value in understanding (as in the presentation of Rosenbaum 2002) why known propensity scores would be useful in an idealized context such as a perfect stratification of the data.

The Utility of Known Propensity Scores. Within a perfect stratification, the propensity score is nothing other than the within-stratum probability of receiving the treatment, or $\Pr(T = 1|S)$ for example 1. In particular, the propensity scores for that example are:

$$\Pr(T = 1|S = 1) = .08/.44 = .182,$$

$$\Pr(T = 1|S = 2) = .12/.24 = .5, \text{ and}$$

$$\Pr(T = 1|S = 3) = .2/.32 = .625.$$

Why is the propensity score useful? As shown earlier for hypothetical examples 1 and 2, if a perfect stratification of the data is available, then the final ingredient for calculating average treatment effect estimates for the treated and for the untreated is the conditional distribution $\Pr(S|T)$. One can recover $\Pr(S|T)$ from the propensity scores by applying Bayes' rule and invoking the marginal distributions of T and S . For example, for the first stratum in Example 1:

$$\Pr(S = 1|T = 1) = \frac{\Pr(T = 1|S = 1)\Pr(S = 1)}{\Pr(T = 1)} = \frac{(.182)(.44)}{(.4)} = .2.$$

Thus, propensity scores encode all of the necessary information about the joint dependence of S and T which is needed to estimate and then combine conditional-on- S treatment effect estimates into estimates of the treatment effect for the treated and the treatment effect for the untreated. Known propensity scores are thus useful for unpacking the inherent heterogeneity of causal effects, and then averaging over such heterogeneity to calculate average treatment effects.

Of course, known propensity scores are almost never available to researchers working with observational rather than experimental data. Thus, the literature on matching more often recognizes the utility of propensity scores for addressing an entirely different concern: solving comparison problems created by the sparseness of data in any finite sample. These methods rely on estimated propensity scores, as we discuss next.

Data Sparseness and Estimated Propensity Scores. Suppose again that a perfect stratification of the data exists and is known. In particular, assumptions A4-S and A5-S are valid for a set of variables in \mathbf{S}_i which can be measured without error. Further suppose that the region of common support extends across the full range of each variable in \mathbf{S}_i , such that the true

propensity score is greater than 0 and less than 1 for every possible combination of values on the variables in \mathbf{S}_i . But, suppose now that (1) there are multiple variables in \mathbf{S}_i , (2) some of these variables take on many values, and (3) the sample is relatively small. In this scenario, there may be many strata in the available data in which no treatment or control cases are observed, even though the true propensity score is between 0 and 1 for every stratum in the population.

Can average treatment effects be consistently estimated in this scenario? Rosenbaum and Rubin (1983a) answer this question affirmatively. The essential points of their argument are the following. First, the sparseness that results from the finiteness of a sample is random, conditional on the joint distribution of \mathbf{S}_i and T . As a result, within each stratum for a perfect stratification of the data, the probability of having a zero-cell in the treatment or the control state is solely a function of the propensity score. And, because such sparseness is conditionally random, strata with identical propensity scores (i.e., different combinations of values for the variables in \mathbf{S}_i but the same within-stratum probability of treatment) can be combined into a more coarse stratification. Over repeated samples from the same population, zero cells would emerge with equal frequency across all strata within these coarse propensity-score-defined strata. Thus, stratifying on the propensity score itself (rather than more finely on \mathbf{S}_i) solves the sparseness problem because the propensity score can be treated as a single perfectly stratifying variable, just as in example 1.

Rosenbaum and Rubin (1983a) argue that if one has observed the variables in \mathbf{S}_i , then the propensity score can be estimated using standard methods, such as logit modeling. That is, one can estimate the propensity score, assuming a logistic distribution:

$$(10) \quad \Pr[T_i = 1 | \mathbf{S}_i] = \frac{\exp(\mathbf{S}_i \boldsymbol{\phi})}{1 + \exp(\mathbf{S}_i \boldsymbol{\phi})}$$

and invoking maximum likelihood to estimate the vector of coefficients in $\boldsymbol{\phi}$. One can then stratify on the index of the estimated propensity score, $e(\mathbf{S}_i) = \mathbf{S}_i \hat{\boldsymbol{\phi}}$, and all of the results established for known propensity scores then obtain.⁸ Consider the following hypothetical example, where reweighting is performed only with respect to the propensity score, resulting in unbiased and consistent estimates of average treatment effects even though sparseness problems are severe.

Hypothetical Example 3

Consider the following Monte Carlo simulation, which is an expanded version of Example 1 in two respects. First, for this example, there are two stratifying variables, S_1 and S_2 , each of which has one hundred separate values. Second, in order to demonstrate the properties of alternative estimators, this example utilizes 50,000 samples of data, each of which is a random realization of the same set of definitions for the constructed variables and the stipulated joint distributions between them.

Generation of the 50,000 Datasets. For the simulation, we gave the variables S_1 and S_2 values of .01, .02, .03, and upward to 1. We then cross-classified the two variables to form a 100 by 100 grid and stipulated a propensity score, as displayed in Figure 1, that is a positive,

⁸ As Rosenbaum (1987) later clarified (see also Rubin and Thomas 1996), the estimated propensity scores do a better job of balancing the observed variables in \mathbf{S}_i than the true propensity scores would in any actual application, since the estimated propensity scores correct for the chance imbalances in \mathbf{S}_i that characterize any finite sample.

non-linear function of both S_1 and S_2 .⁹ We then populated the resulting 20,000 constructed cells (100 by 100 for the S_1 by S_2 grid multiplied by the two values of T) using a Poisson random number generator with the relevant propensity score as the Poisson parameter for the 10,000 cells for the treatment group and one minus the propensity score as the Poisson parameter for the 10,000 cells for the control group. This sampling scheme generates (on average across simulated datasets) the equivalent of 10,000 sample members, assigned to the treatment instead of the control as a function of the probabilities plotted in Figure 1.¹⁰

[INSERT FIGURE 1 ABOUT HERE]

Across the 50,000 simulated datasets, on average 7,728 of the 10,000 possible combinations of values for both S_1 by S_2 had no individuals assigned to the treatment, and 4,813 had no individuals assigned to the control. No matter the actual realized pattern for each simulated dataset, all of the 50,000 datasets are afflicted by substantial sparseness, such that a perfect stratification on all values for the variables S_1 and S_2 would result in many strata within which only treatment or control cases are present.

In order to define treatment effects for each dataset, two potential outcomes were defined as linear functions of individual values S_{1i} and S_{2i} :

$$Y_i^t = 102 + 6S_{1i} + 4S_{2i} + \varepsilon_i^t$$

⁹ The parameterization of Figure 1 is a constrained tensor product spline regression for the index function of a logit. See Ruppert, Wand, and Carroll (2003) for examples of such parameterizations. Figure 1 is generated by setting $\mathbf{S}_i\phi$ in Equation 10 to: $-2+3(S_1)-3(S_1-.1)+2(S_1-.3)-2(S_1-.5)+4(S_1-.7)-4(S_1-.9)+1(S_2)-1(S_2-.1)+2(S_2-.7)-2(S_2-.9)+3(S_1-.5)(S_2-.5)-3(S_1-.7)(S_2-.7)$.

¹⁰ In effect, this set-up establishes S_1 by S_2 as two independent multinomial distributions with equal probability mass for each of their respective 100 values.

$$Y_i^c = 100 + 3S_{1i} + 2S_{2i} + \varepsilon_i^c$$

where both ε_i^t and ε_i^c are independent random draws from a normal distribution with expectation 0 and standard deviation of 5. Then, as in Equation 2, individuals from the treatment group were given an observed Y_i equal to their simulated Y_i^t , and individuals from the control group are given an observed Y_i equal to their simulated Y_i^c .

With this set-up, the simulation makes available 50,000 datasets where the individual treatment effects can be calculated exactly, since true values of Y^t and Y^c are available for all simulated individuals. Since the true average treatment effect, treatment effect for the treated, and treatment effect for the untreated are thus known for each simulated dataset, these average effects can then serve as baselines against which alternative estimators that use data only on Y_i , T_i , S_{1i} and S_{2i} can be compared.

The first row of Table 5 presents true means and standard deviations of the three average treatments effects, calculated across the 50,000 simulated datasets. The mean of the average treatment effect across datasets is 4.525, whereas the means of the average treatment effects for the treated and for the untreated are 4.892 and 4.395, respectively. Similar to hypothetical example 1, this example represents a form of positive selection, where those who are most likely to be in the treatment group are also those most likely to benefit from the treatment. Accordingly, the treatment effect for the treated is larger than the treatment effect for the untreated.

[INSERT TABLE 5 ABOUT HERE]

Methods for Treatment Effect Estimation. Rows 2 through 5 of Table 5 report means and standard deviations across the 50,000 datasets of three sets of regression estimates

of the causal effect of T on Y . The first set is drawn from 50,000 separate regressions of Y on T , resulting in parameter estimates exactly equivalent to what were defined in Equation 4 as naive estimates (because, again, they do not utilize any information about the treatment assignment mechanism). The second and third sets of regression estimates incorporate least squares adjustments for S_1 and S_2 under two different specifications, linear and quadratic.

The last three rows of Table 5 present analogous results for three propensity-score-based weighting estimators. For the estimates in the fifth row, it is (wrongly) assumed that the propensity score can be estimated consistently using a logit model with linear terms for S_1 and S_2 (i.e., assuming that, for Equation 10, a logit with $\mathbf{S}_i\boldsymbol{\phi}$ specified as $\alpha + \phi_1 S_{1i} + \phi_2 S_{2i}$ will yield consistent estimates of the propensity score surface plotted in Figure 1). After the logit model was estimated for each of the 50,000 datasets using the wrong specification, the estimated propensity score for each individual was then calculated:

$$\hat{p}_i = \frac{\exp(\hat{\alpha} + \hat{\phi}_1 S_{1i} + \hat{\phi}_2 S_{2i})}{1 + \exp(\hat{\alpha} + \hat{\phi}_1 S_{1i} + \hat{\phi}_2 S_{2i})}$$

along with the estimated odds of the propensity of being assigned to the treatment:

$$\hat{r}_i = \hat{p}_i / (1 - \hat{p}_i) .^{11}$$

To estimate the treatment effect for the treated, we then implemented a reweighting estimator by calculating the average outcome for the treated and subtracting from this average outcome a counterfactual average outcome using reweighted data on those from the control

¹¹ By definition, all individuals with the same values for S_1 and S_2 have the same estimated propensity scores, and thus it is best to think of each stratum defined by combinations of values of S_1 and S_2 as having an associated propensity score which is then applied to individuals within each stratum, regardless of whether they are in the treatment or control group.

group:

$$\hat{\delta}_{\text{TT},\text{reweight}} \equiv \left\{ \frac{1}{n^t} \sum_{i:T_i=1} Y_i \right\} - \left\{ \frac{\sum_{i:T_i=0} \hat{r}_i Y_i}{\sum_{i:T_i=0} \hat{r}_i} \right\}$$

where n^t is the number of individuals in the treatment group and \hat{r}_i is the estimated odds, as defined earlier, of being in the treatment instead of the control. The reweighting operation in the second term gives more weight to control group individuals equivalent to those in the treatment group, and vice versa (see Rosenbaum 1987, 2002). To estimate the treatment effect for the untreated, we then implemented a reweighting estimator that is the mirror image:

$$\hat{\delta}_{\text{TUT},\text{reweight}} \equiv \left\{ \frac{\sum_{i:T_i=1} Y_i / \hat{r}_i}{\sum_{i:T_i=1} n^t / \hat{r}_i} \right\} - \left\{ \frac{1}{n^c} \sum_{i:T_i=0} Y_i \right\} .$$

where n^c is the number of individuals in the control group. Finally, the corresponding estimator of the unconditional average treatment effect is:

$$\hat{\delta}_{\text{ATE},\text{reweight}} \equiv \left\{ \frac{1}{n} \sum_i T_i \right\} \left\{ \hat{\delta}_{\text{TT},\text{reweight}} \right\} + \left\{ 1 - \frac{1}{n} \sum_i T_i \right\} \left\{ \hat{\delta}_{\text{TUT},\text{reweight}} \right\} ,$$

which is a weighted average of the two conditional average treatment effect estimates.

The same basic reweighting scheme is implemented for the sixth row of Table 5, but the estimated propensity score utilized to define the estimated odds of treatment, \hat{r}_i , is instead based on results from a flawlessly estimated propensity score equation (i.e., one that uses the exact same specification that was fed to the random generator that assigned individuals to the treatment; see note 9 for the tensor product specification). Finally, for the last row of Table 5, the same reweighting scheme is implemented, but, in this case, the estimated odds of

treatment, \hat{r}_i , are replaced by the true odds of treatment, r_i , as calculated with reference to the exact function that generated the propensity score for Figure 1.

Monte Carlo Results. As reported in the second through fourth rows of Table 5, all three regression-based-estimators yield biased and inconsistent estimates of the average treatment effect (which are, on average, too large).¹² As reported in the fifth row of Table 5, the reweighting estimator based on the misspecified logit yields estimates that are closer on average than the regression-based estimators for the average treatment effect. This difference is somewhat artificial, since in general such a difference would depend on the relative misspecification of the propensity score estimating equation, the specification of the alternative regression equation, and the distributions of the potential outcomes.

The sixth row of the Table 5 presents analogous estimates with flawlessly estimated propensity scores. These estimates are unbiased and consistent for the average treatment effect, the treatment effect for the treated, and the treatment effect for the untreated. Finally, for the last row, the reweighting estimates utilize the true propensity scores and are also unbiased and consistent (but, as shown by Rosenbaum (1987, 2002), more variable than those based on the flawlessly estimated propensity score). The last two lines are thus the most important to note, as they demonstrate the claim in the literature: If one can consistently estimate the true propensity score, once can solve entirely the problems created by sparseness of data.

¹² All three regression estimators yield estimates that are typically interpreted as estimates of the average treatment effect, and thus we have placed them in the first column of the table (even though they could be regarded as estimators of other parameters as well, such as the treatment effect for the treated). Notice that, as estimates of the treatment effect for the treated, they are on average too small.

This example shows the potential power of propensity-score-based modeling. If treatment assignment can be modeled perfectly, one can solve the sparseness problems that afflict finite datasets, at least in so far as offering estimates that are unbiased and consistent. On the other hand, this simulation also develops an important qualification of this potential power. Without a perfect specification of the propensity score estimating equation, one cannot rest assured that an unbiased and consistent estimate can be obtained. Since propensity scores achieve their success by undoing the treatment assignment patterns, analogously to reweighting an unrepresentative sample, systematically incorrect estimated propensity scores can generate systematically incorrect weighting schemes that yield biased and inconsistent estimates of treatment effects.

How are Stratification and Reweighting Estimators Forms of Matching?

Given the description of matching estimators offered in the introduction (i.e., algorithms for mechanically identifying matched sets of equivalent treatment and control cases), in what sense are the stratification estimators of hypothetical examples 1 and 2 and then the reweighting estimators of hypothetical example 3 equivalent to matching estimators? Stratification estimators have a straightforward connection. The strata which are formed represent matched sets, and a weighting procedure is then invoked to average stratified treatment effect estimates in order to obtain the average treatment effect of interest. The propensity-score-reweighting-estimators, however, have a less straightforward connection. Here, the data are, in effect, stratified coarsely by the estimation of the propensity score (i.e., since all individual in the same strata, as defined by the stratifying variables in \mathbf{S}_i , are given the same estimated propensity score), and then the reweighting is performed directly across individuals instead of across the strata. This type of

individual-level-reweighting is made necessary because of sparseness (since some of the fine strata for which propensity scores are estimated necessarily contain only treatment or control cases, thereby preventing the direct calculation of stratified treatment effect estimates).

In the opposite direction, it is important to recognize that the algorithmic matching estimators which we summarize in the next section can be considered reweighting estimators. As we show later, these data analysis procedures warrant causal inference by achieving an “as if” stratification of the data that results in a balanced distribution of covariates across matched treated and control cases. Thus, although it is sometimes easier to represent matching estimators as algorithmic data analysis procedures that mechanically match equivalent cases to each other, it is best to understand matching as a method to reweight the data in order to warrant causal inference.¹³

MATCHING ALGORITHMS AS A SET OF DATA ANALYSIS ROUTINES

As shown in hypothetical example 1, if the variables in \mathbf{S}_i are known and observed, then the probability of treatment is random within strata defined by the variables in \mathbf{S}_i (although not random with equal probability in all strata). In this case, treatment assignment is said to be ignorable and the basic matching estimators that we will summarize in this section balance the variables in \mathbf{S}_i across the matched groups of cases selected from the treatment and control

¹³ One might argue that matching can be seen as nothing other than a data analysis procedure with no necessary connection with concerns of causality. Although there is nothing inherently wrong with such a position, it does represent a betrayal of the motivations of more than 60 years of matching. From the 1940s onward, matching estimators have been seen precisely as a method to be utilized when pursuing knowledge about a causal effect of interest (see Greenwood 1945; Freedman and Hawley 1949).

groups. The specific data analysis procedures are straightforward. They are designed to achieve the same outcomes as the reweighting estimators presented earlier in hypothetical example 3, while also allowing one to assess whether or not the supports of the variables in \mathbf{S}_i encompass all treatment and control cases. Thus, in addition to enabling estimation of average treatment effects when all variables in \mathbf{S}_i are known and observed, the methods also facilitate exploratory data analysis to determine whether or not causal effect estimation is even possible.

Variants of Matching Algorithms

Matching estimators differ primarily in (1) the number of matched cases designated for each to-be-matched target case and (2) how multiple matched cases are weighted if more than one is utilized for each target case. In this section, we describe the four main types of matching estimators. All four can be seen as algorithms for implementing a perfect stratification of the data, as presented in idealized scenarios in the last section. But, as we show in the next section, they are data analysis procedures that can be used more generally even when ignorability cannot be assumed (i.e., when some of the variables in \mathbf{S}_i are unobserved), if appropriate care is taken to assess the amount of bias in resulting estimates.

For simplicity of presentation, we will focus on matching estimators of the treatment effect for the treated, although (for this section) all of what we write could be reversed, instead focusing on matching treatment cases to control cases in order to construct an estimate of the treatment effect for the untreated. Heckman et al. (1997, 1998) and Smith and Todd (2005) outline a general notational framework for representing alternative matching estimators. Using our version of potential outcome notation, all matching estimators of the treatment effect for the treated can

be expressed as:

$$(11) \quad \hat{\delta}_{\text{TT,match}} \equiv \frac{1}{n^t} \sum_i \left\{ (Y_i | T_i = 1) - \sum_j \omega_{i,j} (Y_j | T_j = 0) \right\}$$

where n^t is the number of treatment cases, i is the index over treatment cases, j is the index over control cases, and where $\omega_{i,j}$ represents a set of scaled weights that measure the distance between each control case and the target treatment case. Alternative matching estimators can be represented as different procedures for deriving the weights represented by $\omega_{i,j}$. And, thus, in Equation 11, the weights are entirely unspecified. As we will describe next, the weights can take on many values, indeed as many n^t by n^c different values, since alternative weights can be used when constructing the counterfactual for each target treatment case. The difference in the propensity score is the most common distance measure used to construct weights.¹⁴

Exact matching. For the treatment effect for the treated, exact matching constructs the counterfactual for each treatment case using the control cases with identical values on the variables in \mathbf{S}_i . In the notation of Equation 11, exact matching uses weights equal to $1/k$ for matched control cases, where k is the number of matches selected for each target treatment case. Weights of 0 are given to all unmatched control cases. If only one match is chosen randomly from among possible exact matches, then $\omega_{i,j}$ is set to 1 for the randomly selected match (from all available exact matches) and 0 for all other control cases. Exact matching may be combined with any of the matching methods described below by using these methods to match within

¹⁴ Other measures of distance are available, including the difference in the odds of the propensity, the difference in the estimated logit, and the Mahalanobis metric. The Mahalanobis metric is $(\mathbf{S}_i - \mathbf{S}_j)^T \Sigma^{-1} (\mathbf{S}_i - \mathbf{S}_j)$, where Σ is the covariance matrix of the variables in \mathbf{S} (usually calculated for the treatment cases only).

categories of particularly important covariates.

Nearest neighbor matching. For the treatment effect for the treated, nearest neighbor matching constructs the counterfactual for each treatment case using the control cases that are closest to the treatment case on a unidimensional measure constructed from the variables in \mathbf{S}_i , usually an estimated propensity score but sometimes variants of propensity scores (see Althausen and Rubin 1970; Cochran and Rubin 1973; Rubin 1973a, b; Rubin 1976a, b; Rubin 1980; Rosenbaum and Rubin 1983a, 1985). The traditional algorithm randomly orders the treatment cases and then selects for each treatment case the control case with the smallest distance. The algorithm can be run with or without replacement. With replacement, a control case is returned to the pool after a match and can be matched later to another treatment case. Without replacement, a control case is taken out of the pool once it is matched.¹⁵

If only one nearest neighbor is selected for each treatment case, then $\omega_{i,j}$ is set equal to 1 for the matched control case and zero for all other control cases. One can also match multiple nearest neighbors to each target treatment case, in which case $\omega_{i,j}$ is set equal to $1/k_i$ for the matched nearest neighbors, where k_i is the number of matches selected for each target treatment case i . Matching more control cases to each treatment case results in lower expected variance of the treatment effect estimate but also raises the possibility of greater bias, since the probability of making more poor matches increases.

¹⁵ One weakness of the traditional algorithm when used without replacement is that the estimate will vary depending on the initial ordering of the treatment cases. A second weakness is that without replacement the sum distance for all treatment cases will generally not be the minimum because control cases that might make better matches to later treatment cases may be used early in the algorithm. Rosenbaum (1989, 2002) develops an optimal matching algorithm that avoids this problem.

A danger with nearest neighbor matching is that it may result in some very poor matches for treatment cases. A version of nearest neighbor matching, known as caliper matching (or a closely related form known as radius matching [Dehejia and Wahba 2002]), is designed to remedy this drawback by restricting matches to some maximum distance.¹⁶ With this type of matching, some treatment cases may not receive matches, and thus the effect estimate will apply only to the subset of the treatment cases matched (even if ignorability holds and there is simply sparseness in the data).

Interval matching. Interval matching (also referred to as subclassification) sorts the treatment and control cases into segments of a unidimensional metric, usually the estimated propensity score, and then calculates the treatment effect within these intervals (see Cochran 1968; Rosenbaum and Rubin 1983a, 1984; Rubin 1977). For each interval, a variant of the matching estimator in Equation 11 is estimated separately, with $\omega_{i,j}$ chosen to give the same amount of weight to the treatment cases and control cases within each interval. The average treatment effect for the treated is then calculated as the mean of the interval-specific treatment effects, weighted by the number of treatment cases in each interval. This method is nearly indistinguishable from nearest neighbor caliper matching with replacement when each of the intervals includes exactly one treatment case.

Kernel matching. Developed by Heckman et al. (1997, 1998a, b), kernel matching constructs the counterfactual for each treatment case using all control cases but weights each control case based on its distance from the treatment case. Because kernel matching uses more

¹⁶Radius matching involves matching all control cases within a particular distance, the “radius,” from the treatment case and giving the selected control cases equal weight. If there are no control cases within the radius, the nearest available control is used.

cases, it is generally more efficient than other forms of matching, resulting in estimates with smaller expected sampling variance. However, because control cases that are very different from a particular treatment case are used for each treatment case, even though they have a very small weight, there is considerable potential to increase bias. The weights represented by $\omega_{i,j}$ in Equation 11 are calculated using a kernel function, $G(\cdot)$, that transforms the distance between the selected target treatment case and all control cases in the study. When using the estimated propensity score to measure the distance, kernel matching estimators define the weight as:

$$\omega_{i,j} = G\left(\frac{P(S_j) - P(S_i)}{a_n}\right) / \sum_j G\left(\frac{P(S_j) - P(S_i)}{a_n}\right)$$

where a_n is a bandwidth parameter that scales the difference in the estimated propensity scores based on the sample size.¹⁷ The numerator of this expression yields a transformed distance between each control case and the target treatment case. The denominator is a scaling factor equal to the sum of all the transformed distances across control cases, which is needed so that the sum of $\omega_{i,j}$ is equal to 1 across all control cases when matched to each target treatment case.

Although kernel matching estimators appear quite complex, they are a natural extension of interval and nearest neighbor matching: all control cases are matched to each treatment case but weighted so that those closest to the treatment case are given the greatest weight. Smith and Todd (2005) offer an excellent intuitive discussion of kernel matching along with generalizations to local linear matching (Heckman et al. 1997, 1998a, b) and local quadratic matching (Ham et al.

¹⁷ Increasing the bandwidth increases bias but lowers variance. Smith and Todd (2005) find that estimates are fairly insensitive to the size of the bandwidth.

2003).¹⁸

It is sometimes implied in the applied literature that the techniques just summarized are only useful for estimating the treatment effect for the treated. If ignorability can be assumed (that is, all variables in \mathbf{S}_i are known and observed, such that a perfect stratification of the data could be formed with a suitably large dataset) and if the ranges of all of the variables in \mathbf{S}_i are the same for both treatment and control cases, then the matching estimators of this section are consistent for the treatment effect among the treated, the treatment effect among the untreated, and the average treatment effect (just as in hypothetical example 3).

Assessing the Region of Common Support When Implementing a Matching Algorithm

In practice, there is often good reason to believe that some of the lack of observed overlap of \mathbf{S}_i for the treatment and control cases may have emerged from systematic sources, often related to the choice behavior of individuals. In such cases, applied researchers who use matching techniques often focus on the estimation of a treatment effect on the common support of \mathbf{S}_i . Traditionally, this is accomplished by using only treatment/control cases whose propensity scores fall between the minimum and maximum propensity scores in the control/treatment group.

Sometimes matching on the region of common support helps to clarify and sharpen the contribution of a study. When estimating the average treatment effect for the treated, there is no harm in throwing away control cases outside the region of common support if all treatment cases fall within the support of the control cases (i.e., control cases outside the region of common

¹⁸ Rosenbaum (2002, Chapter 10) and Hansen (2004) have developed related forms of full and optimal matching.

support provide no information about the causal effect; recall the cases in which $S = 1$ in hypothetical example 2). And, even if imposing the common support condition results in throwing away some of the treatment cases, this can be considered an important substantive finding, especially for interpreting the treatment effect estimate. In this case, the resulting estimate is the treatment effect for a subset of the treated only, and, in particular, a treatment effect estimate that is informative only about those in the treatment and control groups who are equivalent with respect to observed treatment selection variables. In some applications, this is precisely the estimate needed, for example, when evaluating whether or not a program should be expanded in size in order to accommodate more treatment cases, but without changing eligibility criteria.¹⁹

Coming to terms with these common support issues has become somewhat of a specialized art form within the matching literature. Heckman et al. (1997, 1998a, b) recommend trimming the region of common support to eliminate cases in regions of the common support with extremely low density. This involves selecting a minimum density (labeled the “trimming level”) that is greater than zero. Smith and Todd (2005) note that this solves two potential problems. First, it avoids discarding cases that are just above the maximum or below the minimum but might make good matches. Second, there may be gaps between the minimum and maximum where the supports do not overlap. Heckman et al. (1997) find that estimates are quite sensitive to the level of trimming in small samples, with greater bias when the trimming level is lower. However, increasing the trimming level excludes more treatment cases and results in higher variance.

¹⁹ As argued by Heckman and Vytlačil (1999, 2000, 2004), these types of treatment effect estimates are among the most informative, both for policy guidance and theoretical prediction, as they focus on those at the margin of treatment participation (or causal exposure).

Which Types of Matching Estimators Work Best?

There is very little specific guidance in the literature, and the answer very likely depends on the substantive application. Smith and Todd (2005) and Heckman et al. (1997, 1998a, b) have experimental data against which matching estimators can be compared, and they argue for the advantages of kernel matching (and a particular form of robust kernel matching). To the extent that a general answer to this question can be offered, we would suggest that nearest neighbor caliper matching with replacement, interval matching, and kernel matching are all closely related and should be preferred to nearest neighbor matching without replacement. If the point of a matching estimator is to minimize bias by comparing target cases to similar matched cases, then methods which make it impossible to generate poor matches should be preferred.²⁰

Since there is no clear guidance on which type of matching estimator is “best,” we constructed a fourth hypothetical example to give a sense of how often alternative matching estimators yield appreciably similar estimates. We also develop this example so that it can serve as a bridge to the section that follows, where the substantial additional challenges of real-world applications are discussed.

²⁰ Another criterion for choosing among alternative matching estimators is relative efficiency. Our reading of the literature suggests that little is known about the relative efficiency of these estimators (see especially Abadie and Imbens 2004a, 2004b), even though there are claims in the literature that kernel based methods are the most efficient. The efficiency advantage of kernel matching methods is only a clear guide to practice if kernel-based methods are known to be no more biased than alternatives. But, the relative bias of kernel-based methods are application-dependent, and should interact further with the bandwidth of the kernel. Thus, it seems that we will only know for sure which estimators are most efficient for which types of applications when statisticians discover how to calculate the sampling variances of all alternative estimators. Thereafter, it should be possible to compute mean-squared-error comparisons across alternative estimators for sets of typical applications.

Hypothetical Example 4

For this example, we used simulated data, where we defined the potential outcomes and treatment assignment patterns so that we could explore the relative performance of alternative matching and regression estimators. The former are estimated under alternative scenarios, with three different types of misspecification of the propensity score estimating equation.

Unlike example 3, we do not repeat the simulation for multiple samples but confine ourselves to results on a single sample, as would be typical of any real-world application.

Generation of the Dataset. The dataset that we constructed mimics the dataset from the National Education Longitudinal Study analyzed by Morgan (2001). For that application, Morgan used regression and matching estimators to estimate the effect of Catholic schooling on the achievement of high school students in the United States. For our simulation, we generated a dataset of 10,000 individuals with values for 13 baseline variables that resemble closely the joint distribution of the similar variables in Morgan (2001). The variables include dummies for race, region, urbanicity, have own bedroom, and have two parents, along with an ordinal variable for number of siblings and a continuous variable for socioeconomic status. Then, we created an entirely hypothetical cognitive skill variable, assumed to reflect innate and acquired skills in unknown proportions.²¹

²¹ To be precise, we generated a sample using a multinomial distribution from a race-by-region-by-urbanicity grid from the data in Morgan (2001). We then simulated socioeconomic status as random draws from normal distributions with means and standard deviations estimated separately for each of the race-by-region-by-urbanicity cells using the data from Morgan (2001). Then, we generated all other variables iteratively building on top of these variables, using joint distributions (where possible) based on estimates from the NELS data. Since we relied on standard parametric distributions, the data are somewhat more smooth than the original NELS data (which thereby gives an advantage to parametric regression relative to non-parametric matching methods, as we note later).

We then defined potential outcomes for all 10,000 individuals, assuming that the observed outcome of interest is a standardized test taken at the end of high school. For the potential outcome under the control (i.e., a public school education), we generated “what if” test scores form a normal distribution, with an expectation as a function of race, region, urbanicity, number of siblings, socioeconomic status, family structure, and cognitive skills. We then assumed that the “what if” test scores under the treatment (i.e., a Catholic school education) would be equal to the outcome under the control plus a boosted outcome under the treatment that is function of race, region, and cognitive skills (under the assumption, based on the dominant position in the extant literature, that black and Hispanic respondents from the north, as well as all of those with high cognitive skills, are disproportionately likely to benefit from Catholic schooling).

We then defined the probability of attending a Catholic school using a logit with 26 parameters, based on a specification from Morgan (2001) along with an assumed self-selection dynamic where individuals are slightly more likely to select the treatment as a function of the relative size of their individual-level treatment effect.²² This last component of the logit creates a nearly insurmountable challenge, since in any particular application one would not have such a variable with which to estimate a propensity score. That, however, is

²² The index of the assumed logit was $-4.6 - .69(\text{Asian}) + .23(\text{Hispanic}) - .76(\text{black}) - .46(\text{native American}) + 2.7(\text{urban}) + 1.5(\text{northeast}) + 1.3(\text{north central}) + .35(\text{south}) - .02(\text{siblings}) - .018(\text{bedroom}) + .31(\text{two parents}) + .39(\text{socioeconomic status}) + .33(\text{cognitive skills}) - .032(\text{socioeconomic status squared}) - .23(\text{cognitive skills squared}) - .084(\text{socioeconomic status})(\text{cognitive skills}) - .37(\text{two parents})(\text{black}) + 1.6(\text{northeast})(\text{black}) - .38(\text{north central})(\text{black}) + .72(\text{south})(\text{black}) + .23(\text{two parents})(\text{Hispanic}) - .74(\text{northeast})(\text{Hispanic}) - 1.3(\text{north central})(\text{Hispanic}) - 1.3(\text{south})(\text{Hispanic}) + .25(\text{individual treatment effect} - \text{average treatment effect})$.

our point in including this term, as individuals are thought, in many real-world applications, to be selecting from among alternative treatments based on accurate expectations, unavailable as measures to researchers, of their likely gains from alternative treatment regimes. The probabilities defined by the logit were then passed to a binomial distribution, which resulted in 986 of the 10,000 simulated students attending Catholic schools.

Finally, observed outcomes were assigned according to treatment status. With the sample divided into the treatment group and the control group, we calculated from the pre-specified potential outcomes variables the true baseline average treatment effects. The treatment effect for the treated is 6.96, while the treatment effect for the untreated is 5.9. In combination, the average treatment effect is then 6.0.

Methods for Treatment Effect Estimation. In Table 6, we offer ten separate types of matching estimates. These are based on routines written for STATA by two sets of authors: Becker and Ichino (2002) and Leuven and Sianesi (2003).²³ We estimate all matching estimators under three basic scenarios. First, we offer a set of estimates based on very poorly estimated propensity scores, derived from an estimating equation from which we omitted 9 interaction terms along with the cognitive skill variable. The last specification error is particularly important, as the cognitive skill variable has a correlation of 0.401 with the outcome and 0.110 with the treatment in the simulated data. For the second scenario, we included the cognitive skill variable but continued to omit the 9 interaction terms. For the third scenario, we then added the 9 interaction terms. All three scenarios lack an adjustment

²³ We do not provide a review of software routines, as such a review would be immediately out-of-date upon publication. At present, two additional sets of routines seem to be in use in the applied literature: Ho et al. (2004) and Abadie et al. (2001).

for the self-selection dynamic, in which individuals select into the treatment partly as a function of their expected treatment effect.

[INSERT TABLE 6 ABOUT HERE]

Regarding the specific settings for the alternative matching estimators, the interval matching algorithm began with five blocks and subdivided blocks until each block achieved balance on the estimated propensity score across treatment and control cases (with the final result leading to 10 blocks). Nearest neighbor matching was implemented with replacement and a caliper of 0.002, in both one- and five-nearest-neighbor variants. Radius matching was implemented using a radius of 0.001. For the kernel matching estimators, we used two types of kernels – Epanechnikov and Gaussian – and the default bandwidth of 0.06 for both pieces of software. For the local linear matching estimator, we used the Epanechnikov kernel with the default bandwidth of 0.08. For all matching estimators, we verified that the estimated propensity scores were well balanced across the treatment and control groups, and then that the specific covariates in the respective propensity score estimating equations were similarly well balanced.

Finally, for comparison, we offer, in the last two lines of Table 6, OLS regression estimates of the treatment effect under three analogous scenarios (i.e., including the same variables for the propensity score estimating equation directly in the regression equation). We present regression estimates in two variants: (1) without regard to the distributions of the variables and (2) based on samples restricted to the region of common support (as defined by the propensity score estimated from the covariates utilized for the respective scenario).

Results. We estimated treatment effects under the assumption that self-selection on

the individual Catholic school effect is present, and yet cannot be adjusted for using a statistical model without a measure of individuals' expectations. Thus, we operate under the assumption that only the treatment effect for the treated has any chance of being estimated consistently, as in the study by Morgan (2001) on which this example is based. We therefore compare all estimates to the true treatment effect for the treated, identified earlier as 6.96

Matching estimates using the very poorly estimated propensity scores are reported in the first column of Table 6, along with the implied bias as an estimate of the treatment effect for the treated in the second column (i.e., the matching estimate minus 6.96). As expected, all estimates have a positive bias. Most of the positive bias results from the mistaken exclusion of the cognitive skill variable from the propensity score estimating equation.

Matching estimates using the somewhat poorly estimated propensity scores are reported in the third column of Table 6, along with the implied bias in the fourth column. These estimates are considerably closer to the treatment effect for the treated. Finally, matching estimates using the relatively well estimated propensity scores are reported in the fifth column of Table 6, along with the expected bias in the sixth column. On the whole, these estimates are only slightly better. Fortunately, having the correct specification seems to have reduced the bias in those estimates with the largest bias from column three.²⁴

For comparison, we then provide analogous regression estimates in the final two rows of Table 6. In some cases, these estimates outperform some of the matching estimates. In

²⁴ It is noteworthy that even when we implemented the equivalent matching estimators from both software routines (even beyond those presented in Table 6), we obtained different estimates. We cannot determine the source of these differences from the documentation provided by the software's creators.

fairness to the matching estimates, however, it should be pointed out that the data analyzed for this example are well suited to regression because there are few cases off the region of common support, and the assumed functional form of the potential outcomes is relatively simple.

We have demonstrated two basic points with this example. First, looking across the rows of the table, it is clear that matching estimators and different software routines yield different average treatment effect estimates. Thus, at least for the near future, it will be crucial for researchers to examine multiple estimates of the same treatment effect across estimators and software packages. Second, matching estimators cannot compensate for an unobserved covariate in \mathbf{S} , which leads to comparisons of treatment and control cases that are not identical in all relevant aspects other than treatment status. Even though a matching routine will balance the variables included in the propensity score estimating equation, the resulting matching estimates will remain biased and inconsistent. Unfortunately, violation of the assumption of ignorable treatment assignment is the scenario in which most analysts will find themselves, and this is the scenario to which we turn next.

No matter which type of matching is utilized, it is good practice to examine how much balance on the variables in \mathbf{S}_i across the treatment and control groups has been achieved. As shown by Rosenbaum and Rubin (1984), balance can be tested using a t-test of difference of means across treatment and control cases. If the covariates are not balanced, one can change the estimation model for the propensity score, for example, by adding interaction terms, quadratic terms, or other higher order terms. One can then re-match and re-check the balance. This re-

specification is not considered data mining because it does not involve examining the effect estimate.

Where are the Standard Errors?

Notice that we do not report standard errors for the treatment effect estimates reported in Table 6 for hypothetical example 4. Although there are some simple types of applications in which the variance of matching estimators is known (see Rosenbaum 2002), these are rarely analogous to the situations in which sociologists analyzing observational data will find themselves. Abadie and Imbens (2004a; see also Hahn 1998, Hirano et al. 2003, and Imbens 2003) demonstrate that much remains to be discovered about the sampling variance of alternative estimators. Perhaps most troubling is new work that shows that easy-to-implement bootstrap methods provide correct standard errors of estimates only a very narrow set of circumstances (see Abadie and Imbens 2004b). In the long-run, we suspect that the translation of all matching estimators into forms of semi-parametric regression will continue, and thus that advances following from the work cited here are on the immediate horizon.

MATCHING WHEN TREATMENT ASSIGNMENT IS NON-IGNORABLE

What if the assumption of ignorability of treatment assignment is dubious? That is, what if one only observes a subset of the variables in \mathbf{S}_i , which we will now denote by \mathbf{X}_i . One can still match on \mathbf{X}_i using the techniques just summarized. When in this position, one should (probably) focus first on estimating the treatment effect for the treated. Because a crucial step must be added to the project – assessing or bounding the level of bias resulting from possible non-

ignorability of treatment – focusing on a very specific treatment effect of primary interest helps to ground a discussion of an estimate’s limitations. Then, after using one of the matching estimators of the last section, one should use the data to minimize bias in the estimates, and, if possible, proceed thereafter to a sensitivity analysis. We discuss the possibilities for these steps in the order that analysts usually carry them out.

Covariance adjustment can be incorporated easily into the matching estimators summarized in the last section. Two alternative but similar methods exist. Rubin and Thomas (2000; see also 1996) propose a method that can be used in conjunction with nearest neighbor and interval matching. One simply estimates a regression model on the dataset created by the matching procedure, perhaps re-using some or all of the variables in \mathbf{X}_i , in hopes of relieving unknown consequences of any slight mis-specification of the propensity score estimating equation. The covariates are simply included in the regression model alongside T , generally with fixed effects for alternative strata if multiple cases have been matched to each target case. Heckman et al. (1997; 1998a, b) propose a slightly different procedure. First, one regresses Y on covariates for those in the control group, saving the regression estimates in a vector $\boldsymbol{\beta}^c$. Then, one creates predicted values for all individuals using the variables of particular interest by applying the estimated regression parameters to both the treatment and control cases. Finally, one invokes a matching estimator based on Equation 11, using the residuals in place of outcomes. Abadie and Imbens (2002) show that failure to use a regression adjustment procedure in tandem with a matching algorithm can lead to bias in finite samples in analyses in which \mathbf{S} contains more than one continuous variables. The amount of potential bias increases with the number of variables in the assignment equation. They recommend a simple linear regression adjustment, offering

STATA and MATLAB programs that implements nearest neighbor matching along with the bias correction (see Abadie et al. 2001). Covariance adjustment can also serve to increase the precision of matching estimates (Imbens 2003).

Although these adjustment procedures may help to refine the balance of \mathbf{X}_i across treatment and control cases, they do not help address the problem of unobservable variables in \mathbf{S}_i . These problems can be quite serious if the unobserved variables are fairly subtle, such as a differential latent growth rate for the outcome which is correlated with treatment assignment/selection or an accurate individual-level forecast of the individual-level casual effect. In such cases, the options are quite limited for using the data to diagnose and then correct bias in one's estimates.

If longitudinal data are available, one can incorporate a difference-in-difference adjustment into any of the matching estimators discussed above. For example, when data on the outcome prior to the treatment are available for both the treatment and control cases, one can substitute into Equation 11 the difference between the post-treatment outcome and the pre-treatment outcome for the post-treatment outcome. The difference-in-difference matching estimator controls for all time constant covariates and is analogous to adding individual fixed effects to a regression model. An alternative, as in Dehejia and Wahba (1999), is to include the pre-treatment outcome in the regression equations estimated for the dataset constructed in the matching procedure.

In evaluations of matching estimates of the treatment effect of training programs, Heckman et al. (1997) and Smith and Todd (2005) find that a difference-in-difference local linear matching estimator performed best, coming closest to replicating the experimental estimates of

the effect of the Job Training Partnership Act (JTPA) and National Supported Work (NSW) programs. Whether or not this optimal performance is a reasonable guide for other applications remains to be determined.

Finally, one can perform a sensitivity analysis and/or use the extant literature to discuss the heterogeneity that may lurk beneath the matching estimate. Harding (2003) and DiPrete and Gangl (2004), for example, draw on the tradition of Rosenbaum (1991, 1992, 2002, Rosenbaum and Rubin 1983b) to assess the strength of the relationship that an unobserved variable would have to have with a treatment and an outcome variable in order to challenge a causal inference. Morgan (2001) analyzes variation in the treatment effect estimate across quintiles of the estimated propensity score, offering alternative interpretations of variation in treatment effect estimates based on competing positions in the relevant applied literature about the nature of some crucial unobserved variables.

MATCHING AND REGRESSION

When matching estimators are seen as alternatives to regression estimators of causal effects, a simple question invariably arises: When should matching be used instead of (and to the exclusion of) regression? We view this question as artificial, as matching and regression are closely related, and each can be seen as a variant of the other.

Regression can be used as a technique to execute a stratification of the data. For hypothetical example 1, one could specify S as two dummy variables and T as one dummy variable. If all two-way interactions between S and T are then included in a regression model predicting the outcome, then one has enacted the perfect stratification by fitting a saturated model

to the cells of the first panel of Table 2. Accordingly, if one obtains the marginal distribution of S and the joint distribution of S given T , then one can properly average the coefficient contrasts across the relevant distributions of S in order to obtain consistent estimates of the average treatment effect, the treatment effect among the treated, and the treatment effect among the untreated.

More generally, the relationship between matching and regression has been established in the recent econometrics literature. Kernel matching is equivalent to some varieties of semi-parametric regression (Hahn 1998; Hirano et al. 2003), and least squares regression can be seen as a variance-weighted form of interval matching (Angrist and Krueger 1999). Moreover, all average causal effect estimators can be interpreted as weighted averages of marginal treatment effects (Heckman and Vytlačil 2004), whether generated by matching, regression, or local instrumental variable estimators.

Nonetheless, regression can yield misleading results. If, for hypothetical example 1, S were entered as a simple linear term interacted with T (or instead if S were entered as two dummy variables but not interacted with T), regression would yield coefficient contrasts that mask the underlying treatment effects. Rubin (1977) provides simple and elegant examples of all such complications, highlighting the importance of parallel response surface assumptions for regression estimators (see also Holland and Rubin 1983 and Rosenbaum 1984).

Moreover, regression can mask underlying support problems. Consider hypothetical example 2 depicted in Table 3. If the saturated model is fit to the data, the lack of overlapping support will be revealed to the analyst, as the regression routine will drop the coefficient for the zero cell. However, if a constrained version of the model were fit, such as if S were entered as a

simple linear term interacted with T , the regression model would yield seemingly reasonable coefficients. In cases such as this one, regression modeling makes it too easy for an analyst to overlook the potential complications of support conditions. And, thus, one can obtain average treatment effect estimates even when no meaningful average treatment effect exists.

CONCLUSIONS

Matching focuses attention on the heterogeneity of the causal effect. It forces the analyst to examine the alternative distributions of covariates across those exposed to different levels of the causal variable. In the process, it helps the analyst to recognize which cases in the study are incomparable, such as which control cases can be ignored when estimating the treatment effect for the treated and which treatment cases may have no meaningful counterparts among the controls. Finally, matching helps to motivate more sophisticated discussions of the unobservables that may be correlated with the causal variable, and this is an advance over merely conceding that selection bias may be present in some form and speculating on the sign of the bias. Thus, although matching does not solve all (or even very many) of the problems that prevent regression models from generating reliable estimates of causal effects, matching succeeds admirably in laying bare the particular problems of estimating causal effects and then motivating the future research that is needed to resolve causal controversies.

Although these are the advantages of matching, it is important that we not oversell the potential power of the techniques. In much of the applied literature on matching, the propensity score is presented as a single predictive dimension that can be used to balance the distribution of important covariates across treatment and control cases, thereby warranting causal inference. As

we showed in hypothetical example 4, perfect balance on important covariates does not necessarily warrant causal claims. If one does not know of variables that, in an infinite sample, would yield a perfect stratification, then simply predicting treatments status from the observed variables using a logit model and then matching on the estimated propensity score does not solve the causal inference problem. The estimated propensity scores will balance those variables across the treatment and control cases. But, the study will remain open to the sort of “hidden bias” explored by Rosenbaum (2002) but which is often labeled selection on the unobservables in the social sciences. Matching, like regression, is thus a statistical method for analyzing available data, which may have some advantages in some situations. But, in the end, matching cannot compensate for data insufficiency. Causal controversies are best resolved by collecting new and better data.

REFERENCES

- Abadie, Alberto, David Drukker, Jane L. Herr, and Guido W. Imbens. 2001. "Implementing Matching Estimators for Average Treatment Effects in Stata." *The Stata Journal* 1:1-18.
- Abadie, Alberto and Guido W. Imbens. 2004a. "Large Sample Properties of Matching Estimators for Average Treatment Effects." Working Paper, John F. Kennedy School of Government, Harvard University.
- . 2004b. "On the Failure of the Bootstrap for Matching Estimators." Working Paper, John F. Kennedy School of Government, Harvard University.
- Althausen, Robert P. and Donald B. Rubin. 1970. "The Computerized Construction of a Matched Sample." *American Journal of Sociology* 76:325-46.
- . 1971. "Measurement Error and Regression to the Mean in Matched Samples." *Social Forces* 50:206-14.
- Angrist, Joshua D. and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." Pp. 1277-1366 in *Handbook of Labor Economics*, vol. 3, edited by O. C. Ashenfelter and D. Card. Amsterdam: Elsevier.
- Becker, Sascha O. and Andrea Ichino. 2002. "Estimation of Average Treatment Effects Based on Propensity Scores." *The Stata Journal* 2:358-77.
- Berk, Richard A. and Phyllis J. Newton. 1985. "Does Arrest Really Deter Wife Battery? An Effort to Replicate the Findings of the Minneapolis Spouse Abuse Experiment." *American Sociological Review* 50:253-62.
- Berk, Richard A., Phyllis J. Newton, and Sarah Fenstermaker Berk. 1986. "What a Difference a Day Makes: An Empirical Study of the Impact of Shelters for Battered Women." *Journal of Marriage and the Family* 48:481-90.
- Cochran, W.G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24:295-313.
- Dehejia, Rajeev H. and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*:1053-62.
- . 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84:151-61.
- DiPrete, Thomas A. and Henriette Engelhardt. 2004. "Estimating Causal Effects with Matching Methods in the Presence and Absence of Bias Cancellation." *Sociological Methods and Research* 32:501-28.
- DiPrete, Thomas A. and Markus Gangl. 2004. "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments." *Sociological Methodology* 34:271-310.
- Freedman, Ronald and Amos H. Hawley. 1949. "Unemployment and Migration in the Depression." *Journal of the American Statistical Association* 44:260-72.
- Greenwood, Ernest. 1945. *Experimental Sociology: A Study in Method*. New York: King's Crown Press.
- Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66:315-31.

- Ham, J. C., X. Li, and P. B. Reagan. 2003. "Propensity Score Matching, a Distance-Based Measure of Migration, and the Wage Growth of Young Men. Working Paper, Department of Sociology and Center for Human Resource Research, the Ohio State University."
- Hansen, Ben B. 2004. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association* 99:609-18.
- Harding, David J. 2003. "Counterfactual Models of Neighborhood Effects: The Effect of Neighborhood Poverty on Dropping out and Teenage Pregnancy." *American Journal of Sociology* 109:676-719.
- Heckman, James J. 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." *The Quarterly Journal of Economics* 115:45-97.
- Heckman, James J., Hidehiko Ichimura, Jeffery A. Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66:1017-98.
- Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64:605-54.
- Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65:261-94.
- Heckman, James J., Robert J. LaLonde, and Jeffery A. Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." Pp. 1865-2097 in *Handbook of Labor Economics*, vol. 3, edited by O. C. Ashenfelter and D. Card. Amsterdam: Elsevier.
- Heckman, James J., Justin L. Tobias, and Edward Vytlacil. 2003. "Simple Estimators for Treatment Parameters in a Latent-Variable Framework." *The Review of Economics and Statistics* 85:748-55.
- Heckman, James J. and Edward Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences of the United States of America* 96:4730-34.
- . 2000. "The Relationship between Treatment Parameters within a Latent Variable Framework." *Economics Letters* 66:33-39.
- . 2004. "Structural Equations, Treatment Effects and Econometric Policy Evaluation." *Econometrica*.
- Hirano, Keisuke and Guido W. Imbens. 2004. "The Propensity Score with Continuous Treatments." Pp. 73-84 in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, edited by A. Gelman and X.-L. Meng. New York: Wiley.
- Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71:1161-89.
- Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2004. "Matchit." (See <http://gking.harvard.edu/matchit/>).
- Hoffer, Thomas, Andrew M. Greeley, and James S. Coleman. 1985. "Achievement Growth in Public and Catholic Schools." *Sociology of Education* 58:74-97.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945-70.

- Holland, Paul W. and Donald B. Rubin. 1983. "On Lord's Paradox." Pp. 3-25 in *Principles of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*, edited by H. Wainer and S. Messick. Hillsdale: Erlbaum.
- Imai, Kosuke and David A. van Dyk. 2004. "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99.
- Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87:706-10.
- . 2003. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." NBER Technical Working Paper No. 294, National Bureau of Economic Research, Cambridge, Mass.
- Lechner, Michael. 2002a. "Some Practical Issues in the Evaluation of Heterogeneous Labour Market Programmes by Matching Methods." *Journal of Royal Statistical Society* 165:59-82.
- . 2002b. "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies." *The Review of Economics and Statistics* 84:205-20.
- Leuven, Edwin and Barbara Sianesi. 2003. "Psmatch2."
- Lu, Bo, Elaine Zanutto, Robert Hornik, and Paul R. Rosenbaum. 2001. "Matching with Doses in an Observational Study of a Media Campaign against Drug Abuse." *Journal of the American Statistical Association* 96:1245.
- Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.
- Morgan, Stephen L. 2001. "Counterfactuals, Causal Effect Heterogeneity, and the Catholic School Effect on Learning." *Sociology of Education* 74:341-74.
- Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.
- Rosenbaum, Paul R. 1984. "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society, Series A* 147:656-66.
- . 1987. "Model-Based Direct Adjustment." *Journal of the American Statistical Association* 82:387-94.
- . 1989. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association* 84:1024-32.
- . 1991. "Sensitivity Analysis for Matched Case Control Studies." *Biometrics* 47:87-100.
- . 1992. "Detecting Bias with Confidence in Observational Studies." *Biometrika* 79:367-74.
- . 2002. *Observational Studies*. New York: Springer.
- Rosenbaum, Paul R. and Donald B. Rubin. 1983a. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41-55.
- . 1983b. "Assessing Sensitivity to an Unobserved Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society* 45:212-18.
- . 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79:516-24.
- . 1985a. "Constructing a Control Group Using Multivariate Matched Sampling Methods." *The*

- American Statistician* 39:33-8.
- . 1985b. "The Bias Due to Incomplete Matching." *Biometrics* 41:103-16.
- Rubin, Donald. B. 1973a. "Matching to Remove Bias in Observational Studies." *Biometrics* 29:159-83.
- . 1973b. "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies." *Biometrics* 29:185-203.
- . 1976a. "Multivariate Matching Methods That Are Equal Percent Bias Reducing, I: Some Examples." *Biometrics* 32:109-20.
- . 1976b. "Multivariate Matching Methods That Are Equal Percent Bias Reducing, Ii: Maximums on Bias Reduction for Fixed Sample Sizes." *Biometrics* 32:121-32.
- . 1977. "Assignment to Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2:1-26.
- . 1979. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74:318-28.
- . 1980. "Bias Reduction Using Mahalanobis-Metric Matching." *Biometrics* 36:293-98.
- Rubin, Donald B. and Neal Thomas. 1996. "Matching Using Estimated Propensity Scores: Relating Theory to Practice." *Biometrics* 52:249-64.
- . 2000. "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates." *Journal of the American Statistical Association* 95:573-85.
- Ruppert, David, M. P. Wand, and Raymond J. Carroll. 2003. *Semiparametric Regression*. Cambridge: Cambridge University Press.
- Smith, Herbert L. 1997. "Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology* 27:325-53.
- Smith, Jeffery A. and Petra Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125:305-53.
- Sobel, Michael E. 1995. "Causal Inference in the Social and Behavioral Sciences." Pp. 1-38 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by G. Arminger, C. C. Clogg, and M. E. Sobel. New York: Plenum.
- Winship, Christopher and Stephen L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *Annual Review of Sociology* 25:659-706.
- Yinger, Milton J., Kiyoshi Ikeda, and Frank Laycock. 1967. "Treating Matching as a Variable in a Sociological Experiment." *American Sociological Review* 32:801-12.

Table 1. The Joint Probability Distribution and Average Potential Outcomes for Hypothetical Example 1

Joint Probability Distribution of S and T			
	$T = 0$	$T = 1$	
$S = 1$	$\Pr[S = 1, T = 0] = .36$	$\Pr[S = 1, T = 1] = .08$	$\Pr[S = 1] = .44$
$S = 2$	$\Pr[S = 2, T = 0] = .12$	$\Pr[S = 2, T = 1] = .12$	$\Pr[S = 2] = .24$
$S = 3$	$\Pr[S = 3, T = 0] = .12$	$\Pr[S = 3, T = 1] = .2$	$\Pr[S = 3] = .32$
	$\Pr[T = 0] = .6$	$\Pr[T = 1] = .4$	
Potential Outcomes			
	Under the Control State	Under the Treatment State	
$S = 1$	$E[Y^c S = 1] = 2$	$E[Y^t S = 1] = 4$	$E[Y^t - Y^c S = 1] = 2$
$S = 2$	$E[Y^c S = 2] = 6$	$E[Y^t S = 2] = 8$	$E[Y^t - Y^c S = 2] = 2$
$S = 3$	$E[Y^c S = 3] = 10$	$E[Y^t S = 3] = 14$	$E[Y^t - Y^c S = 3] = 4$
	$E[Y^c T = 0] = (.36/.6)(2) + (.12/.6)(6) + (.12/.6)10$ $= 4.4$	$E[Y^t T = 1] = (.08/.4)(4) + (.12/.4)(8) + (.2/.4)14$ $= 10.2$	

Table 2. Estimated Conditional Expectations and Probabilities from a Very Large Sample for Hypothetical Example 1

Estimated Observable Outcomes Conditional on the Stratifying Variable S and the Treatment Variable T		
	Control Group	Treatment Group
$S = 1$	$E_N[Y_i T_i = 0, S_i = 1] = 2$	$E_N[Y_i T_i = 1, S_i = 1] = 4$
$S = 2$	$E_N[Y_i T_i = 0, S_i = 2] = 6$	$E_N[Y_i T_i = 1, S_i = 2] = 8$
$S = 3$	$E_N[Y_i T_i = 0, S_i = 3] = 10$	$E_N[Y_i T_i = 1, S_i = 3] = 14$
Estimated Conditional Distribution of the Stratifying Variable S given the Treatment Variable T		
$S = 1$	$\Pr_N[S_i = 1 T_i = 0] = .6$	$\Pr_N[S_i = 1 T_i = 1] = .2$
$S = 2$	$\Pr_N[S_i = 2 T_i = 0] = .2$	$\Pr_N[S_i = 2 T_i = 1] = .3$
$S = 3$	$\Pr_N[S_i = 3 T_i = 0] = .2$	$\Pr_N[S_i = 3 T_i = 1] = .5$

Table 3. The Joint Probability Distribution and Average Potential Outcomes for Hypothetical Example 2

Joint Probability Distribution of S and T			
	$T = 0$	$T = 1$	
$S = 1$	$\Pr[S = 1, T = 0] = .4$	$\Pr[S = 1, T = 1] = 0$	$\Pr[S = 1] = .4$
$S = 2$	$\Pr[S = 2, T = 0] = .1$	$\Pr[S = 2, T = 1] = .13$	$\Pr[S = 2] = .23$
$S = 3$	$\Pr[S = 3, T = 0] = .1$	$\Pr[S = 3, T = 1] = .27$	$\Pr[S = 3] = .37$
	$\Pr[T = 0] = .6$	$\Pr[T = 1] = .4$	
Potential Outcomes			
	Under the Control State	Under the Treatment State	
$S = 1$	$E[Y^c S = 1] = 2$		
$S = 2$	$E[Y^c S = 2] = 6$	$E[Y^t S = 2] = 8$	$E[Y^t - Y^c S = 2] = 2$
$S = 3$	$E[Y^c S = 3] = 10$	$E[Y^t S = 3] = 14$	$E[Y^t - Y^c S = 3] = 4$
	$E[Y^c T = 0] = (.4/.6)(2) + (.1/.6)(6) + (.1/.6)10$ $= 4$	$E[Y^t T = 1] = (.13/.4)(8) + (.27/.4)14$ $= 12.05$	

Table 4. Estimated Conditional Expectations and Probabilities from a Very Large Sample for Hypothetical Example 2

Estimated Observable Outcomes Conditional on the Stratifying Variable S and the Treatment Variable T		
	Control Group	Treatment Group
$S = 1$	$E_N[Y_i T_i = 0, S_i = 1] = 2$	
$S = 2$	$E_N[Y_i T_i = 0, S_i = 2] = 6$	$E_N[Y_i T_i = 1, S_i = 2] = 8$
$S = 3$	$E_N[Y_i T_i = 0, S_i = 3] = 10$	$E_N[Y_i T_i = 1, S_i = 3] = 14$
Estimated Conditional Distribution of the Stratifying Variable S given the Treatment Variable T		
$S = 1$	$\Pr_N[S_i = 1 T_i = 0] = .667$	$\Pr_N[S_i = 1 T_i = 1] = 0$
$S = 2$	$\Pr_N[S_i = 2 T_i = 0] = .167$	$\Pr_N[S_i = 2 T_i = 1] = .325$
$S = 3$	$\Pr_N[S_i = 3 T_i = 0] = .167$	$\Pr_N[S_i = 3 T_i = 1] = .675$

Table 5. Monte Carlo Means and Standard Deviations of Average Treatment Effects for Hypothetical Example 3

	Average Treatment Effect	Average Treatment Effect for the Treated	Average Treatment Effect for the Untreated
Baseline True Effects			
Simple averages of individual-level treatment effects	4.525 (.071)	4.892 (.139)	4.395 (.083)
Regression Estimators			
OLS estimates from the regression of Y on T (i.e., the naive estimator)	5.388 (.121)		
OLS estimates from the regression of Y on T , S_1 , and S_2	4.753 (.117)		
OLS estimates from the Regression of Y on T , S_1 , S_1 squared, S_2 , and S_2 squared	4.739 (.118)		
Propensity-Score-Based Estimators			
Reweighting estimates based on a misspecified set of propensity score estimates	4.456 (.122)	4.913 (.119)	4.293 (.128)
Reweighting estimates based on a perfectly specified set of propensity score estimates	4.526 (.120)	4.892 (.121)	4.396 (.125)
Reweighting estimates based on the true propensity scores	4.527 (.127)	4.892 (.127)	4.396 (.132)

Notes: Monte Carlo standard deviations, across the 50,000 simulated datasets, are in parentheses.

Table 6. Matching and Regression Estimates for the Simulated Effect of Catholic Schooling on Achievement, as Specified for Hypothetical Example 4

	Very Poorly Specified Propensity Score Estimating Equation (No cognitive skill variable utilized, missing interaction terms, and no adjustment for self-selection on the causal effect)		Poorly Specified Propensity Score Estimating Equation (Missing interaction terms and no adjustment for self-selection on the causal effect)		Well Specified Propensity Score Estimating Equation (No adjustment for self-selection on the causal effect)	
	Treatment Effect Estimate	Bias as an estimate of the treatment effect for the treated	Treatment Effect Estimate	Bias as an estimate of the treatment effect for the treated	Treatment Effect Estimate	Bias as an estimate of the treatment effect for the treated
Matching						
Interval with 10 blocks (B&I)	7.93	0.97	6.83	-0.13	6.73	-0.23
1 Nearest Neighbor with caliper = 0.001 (L&S)	8.16	1.20	6.45	-0.51	6.69	-0.27
5 Nearest Neighbors with caliper = 0.001 (L&S)	7.97	1.01	6.77	-0.19	7.04	0.08
Radius with radius = 0.001 (L&S)	8.02	1.06	6.73	-0.23	6.90	-0.06
Radius with radius = 0.001 (B&I)	8.13	1.17	7.55	0.59	7.29	0.33
Kernel with Epanechnikov kernel (L&S)	7.97	1.01	7.09	0.13	6.96	0.00
Kernel with Epanechnikov kernel (B&I)	7.89	0.93	6.99	0.03	6.86	-0.10
Kernel with Gaussian kernel (L&S)	8.09	1.13	7.31	0.35	7.18	0.22
Kernel with Gaussian kernel (B&I)	7.97	1.01	7.15	0.19	7.03	0.09
Local Linear with Epanechnikov kernel (L&S)	7.91	7.91	6.83	6.83	6.84	-0.12
OLS Regression						
Not Restricted to the Region of Common Support	7.79	0.83	6.84	-0.12	6.81	-0.15
Restricted to the Region of Common Support	7.88	0.92	6.91	-0.05	6.80	-0.16

Notes: All matching estimates on region of common support; B&I = Becker and Ichino software; L&S = Leuven & Sianesi software.

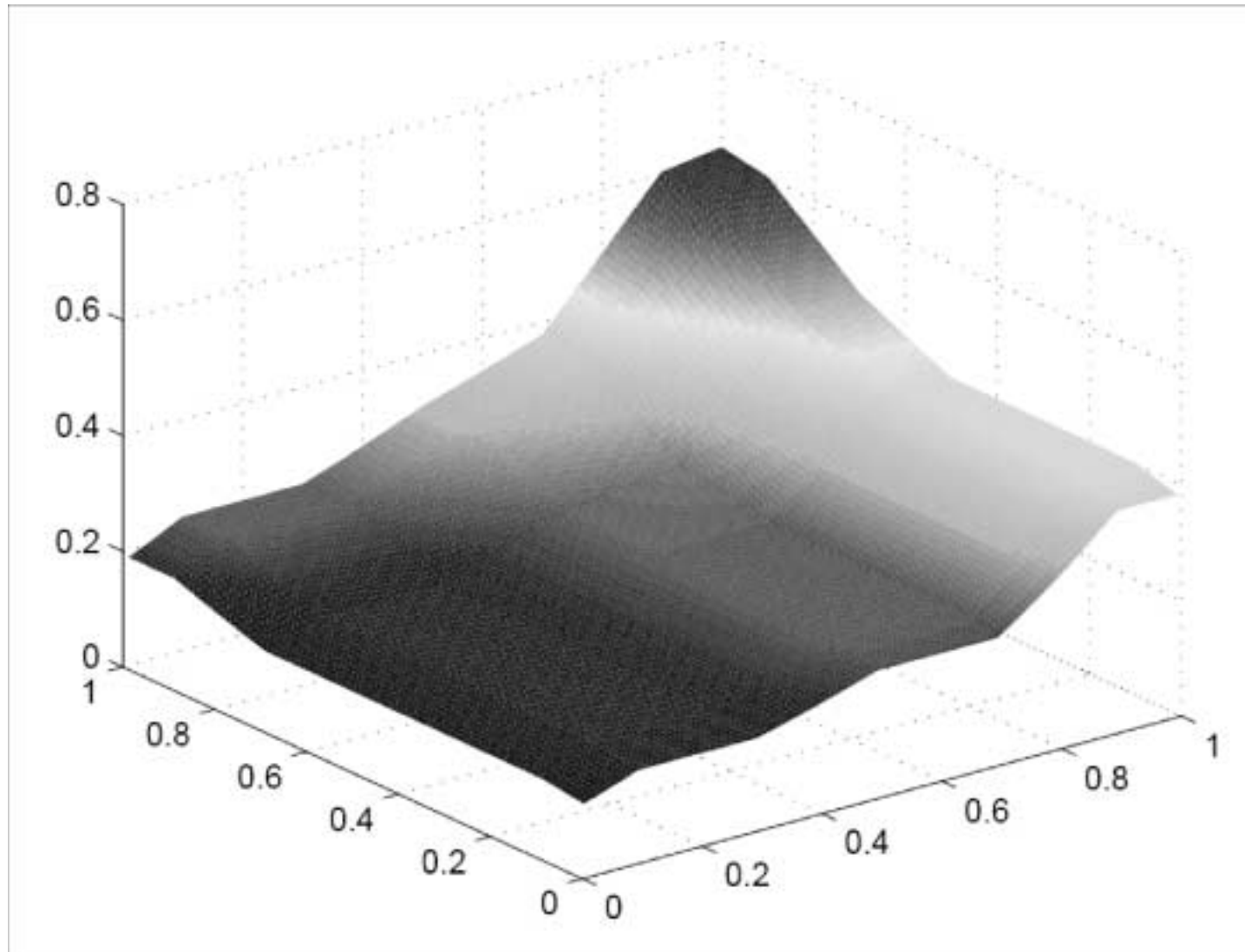


Figure 1. The True Propensity for Hypothetical Example 3 as a Function of S_1 and S_2