Opiates for the Matches: Matching Methods for Causal Inference

Jasjeet S. Sekhon

Travers Department of Political Science, Survey Research Center, University of California, Berkeley, California 94720; email: sekhon@berkeley.edu

Annu. Rev. Polit. Sci. 2009. 12:487–508

The *Annual Review of Political Science* is online at polisci.annualreviews.org

This article's doi:
10.1146/annurev.polisci.11.060606.135444

1094-2939/09/0615-0487$20.00

Key Words

causal inference, matching, Neyman-Rubin model

Abstract

In recent years, there has been a burst of innovative work on methods for estimating causal effects using observational data. Much of this work has extended and brought a renewed focus on old approaches such as matching, which is the focus of this review. The new developments highlight an old tension in the social sciences: a focus on research design versus a focus on quantitative models. This realization, along with the renewed interest in field experiments, has marked the return of foundational questions as opposed to a fascination with the latest estimator. I use studies of get-out-the-vote interventions to exemplify this development. Without an experiment, natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive.

INTRODUCTION

RD: regression discontinuity

Although the quantitative turn in the search for causal inferences is more than a century old in the social sciences, in recent years there has been a renewed interest in the problems associated with making causal inferences using such methods. These recent developments highlight tensions in the quantitative tradition that have been present from the beginning. There are a number of conflicting approaches, which overlap but have important distinctions. I focus here on three of them: the experimental, the model-based, and the design-based.

The first is the use of randomized experiments, which in political science may go back to Gosnell (1927).[1] Whether Gosnell randomized or not, Eldersveld (1956) certainly did when he conducted a randomized field experiment to study the effectiveness of canvassing by mail, telephone, and house-to-house visits on voter mobilization. But even with randomization, there is ample disagreement and confusion about exactly how such data should be analyzed—for example, is adjustment by multivariate regression unbiased? There are also concerns about external validity and whether experiments can be used to answer "interesting" or "important" questions. This latter concern appears to be common among social scientists and is sometimes harshly put. One early and suspicious reviewer of experimental methods in the social sciences recalled the words of Horace: "*Parturiunt montes, nascetur ridiculus mus*" (Mueller 1945).[2] For observational data analysis, however, the disagreements are sharper.

By far the dominant method of making causal inferences in the quantitative social sciences is model-based, and the most popular model is multivariate regression. This tradition is also surprisingly old; the first use of regression to estimate treatment effects (as opposed to simply fitting a line through data) was Yule's (1899) investigation into the causes of changes in pauperism in England. By that time the understanding of regression had evolved from what Stigler (1990) calls the Gauss-Laplace synthesis. The third tradition focuses on design. Examples abound, but they can be broadly categorized as natural experiments or regression-discontinuity (RD) designs. They share in common an assumption that found data, not part of an actual field experiment, have some "as if random" component: that the assignment to treatment can be regarded as if it were random, or can be so treated after some covariate adjustment. From the beginning, some natural experiments were analyzed as if they were actual experiments (e.g., difference of means), others by matching methods (e.g., Chapin 1938), and yet others—many, many others—by instrumental variables (e.g., Yule 1899). [For an interesting note on who invented instrumental variable regression, see Stock & Trebbi (2003).] A central criticism of natural experiments is that they are not randomized experiments. In most cases, the "as if random" assumption is implausible (for reviews see Dunning 2008 and Rosenzweig & Wolpin 2000).

Regression-discontinuity was first proposed by Thistlethwaite & Campbell (1960). They proposed RD as an alternative to what they called "ex post facto experiments," or what we today would call natural experiments analyzed by matching methods. More specifically, they proposed RD as an alternative to matching methods and other "as if" (conditionally) random experiments outlined by Chapin (1938) and Greenwood (1945), where the assignment mechanism is not well understood. In the case of RD, the researcher finds a sharp breakpoint that makes seemingly random distinctions between units that receive treatment and those that do not.

[1] Gosnell may not have actually used randomization (Green & Gerber 2002). His 1924 get-out-the-vote experiment, described in his 1927 book, was conducted one year before Fisher's 1925 book and 11 years before Fisher's famous 1935 book on experimental design. Therefore, unsurprisingly, Gosnell's terminology is nonstandard and leads to some uncertainty about exactly what was done. A definitive answer requires a close examination of Gosnell's papers at the University of Chicago.

[2] "The mountains are in labor, a ridiculous mouse will be brought forth," from Horace's, *Epistles*, Book II, *Ars Poetica* (The Art of Poetry). Horace is observing that some poets make great promises that result in little.

Where does matching fit in? As we shall see, it depends on how it is used.

One of the innovative intellectual developments over the past few years has been to unify all of these methods into a common mathematical and conceptual language, that of the Neyman-Rubin model (Neyman 1990 [1923], Rubin 1974). Although randomized experiments and matching estimators have long been tied to the model, recently instrumental variables (Angrist et al. 1996) and RD (Lee 2008) have also been so tied. This leads to an interesting unity of thought that makes clear that the Neyman-Rubin model is the core of the causal enterprise, and that the various methods and estimators consistent with it, although practically important, are of secondary interest. These are fighting words, because all of these techniques, particularly the clearly algorithmic ones such as matching, can be used without any ties to the Neyman-Rubin model or causality. In such cases, matching becomes nothing more than a nonparametric estimator, a method to be considered alongside CART (Breiman et al. 1984), BART (Chipman et al. 2006), kernel estimation, and a host of others. Matching becomes simply a way to lessen model dependence, not a method for estimating causal effects per se. For causal inference, issues of design are of utmost importance; a lot more is needed than just an algorithm. Like other methods, matching algorithms can always be used, and they usually are, even when design issues are ignored in order to obtain a nonparametric estimate from the data. Of course, in such cases, what exactly has been estimated is unclear.

The Neyman-Rubin model has radical implications for work in the social sciences given current practices. According to this framework, much of the quantitative work that claims to be causal is not well posed. The questions asked are too vague, and the design is hopelessly compromised by, for example, conditioning on post-treatment variables (Cox 1958, Section 4.2; Rosenbaum 2002, pp. 73–74).

The radical import of the Neyman-Rubin model may be highlighted by using it to determine how regression estimators behave when fitted to data from randomized experiments. Randomization does not justify the regression assumptions (Freedman 2008b,c). Without additional assumptions, multiple regression is not unbiased. The variance estimates from multiple regression may be arbitrarily too large or too small, even asymptotically. And for logistic regression, matters only become worse (Freedman 2008d). These are fearful conclusions. These pathologies occur even with randomization, which is supposed to be the easy case.

Although the Neyman-Rubin model is currently the most prominent, and I focus on it in this review, there have obviously been many other attempts to understand causal inference (reviewed by Brady 2008). An alternative whose prominence has been growing in recent years is Pearl's (2000) work on nonparametric structural equations models (for a critique see Freedman 2004). Pearl's approach is a modern reincarnation of an old enterprise that has a rich history, including foundational work on causality in systems of structural equations by the political scientist Herbert Simon (1953). Haavelmo (1943) was the first to precisely examine issues of causality in the context of linear structural equations with random errors.

As for matching itself, there is no consensus on how exactly matching ought to be done, how to measure the success of the matching procedure, and whether or not matching estimators are sufficiently robust to misspecification so as to be useful in practice (Heckman et al. 1998). To illuminate issues of general interest, I review a prominent exchange in the political science literature involving a set of get-out-the-vote (GOTV) field experiments and the use of matching estimators (Arceneaux et al. 2006; Gerber & Green 2000, 2005; Hansen & Bowers 2009; Imai 2005).

The matching literature is growing rapidly, so it is impossible to summarize it in a brief review. I focus on design issues more than the technical details of exactly how matching should be done, although the basics are reviewed. Imbens & Wooldridge (2008) have provided an excellent review of recent

developments in methods for program evaluation. For additional reviews of the matching literature, see Morgan & Harding (2006), Morgan & Winship (2007), Rosenbaum (2005), and Rubin (2006).

THE NEYMAN-RUBIN CAUSAL MODEL

The Neyman-Rubin framework has become increasingly popular in many fields, including statistics (Holland 1986; Rosenbaum 2002; Rubin 1974, 2006), medicine (Christakis & Iwashyna 2003, Rubin 1997), economics (Abadie & Imbens 2006a; Dehejia & Wahba 2002, 1999; Galiani et al. 2005), political science (Bowers & Hansen 2005, Imai 2005, Sekhon 2004), sociology (Diprete & Engelhardt 2004, Morgan & Harding 2006, Smith 1997, Winship & Morgan 1999), and even law (Rubin 2001). The framework originated with Neyman's (1990 [1923]) model, which is nonparametric for a finite number of treatments where each unit has two potential outcomes for each treatment—one if the unit is treated and the other if untreated. A causal effect is defined as the difference between the two potential outcomes, but only one of the two potential outcomes is observed. Rubin (1974, 2006) developed the model into a general framework for causal inference with implications for observational research. Holland (1986) wrote an influential review article that highlighted some of the philosophical implications of the framework. Consequently, instead of the "Neyman-Rubin model," the model is often simply called the Rubin causal model (e.g., Holland 1986) or sometimes the Neyman-Rubin-Holland model (e.g., Brady 2008) or the Neyman-Holland-Rubin model (e.g., Freedman 2006).

The intellectual history of the Neyman-Rubin model is the subject of some controversy (e.g., Freedman 2006, Rubin 1990, Speed 1990). Neyman's 1923 article never mentions the random assignment of treatments. Instead, the original motivation was an urn model, and the explicit suggestion to use the urn model to physically assign treatments is absent from the paper (Speed 1990). An urn model is based on an idealized thought experiment in which colored balls are drawn randomly from an urn. Using the model does not imply that treatment should be physically assigned in a random fashion. It was left to R.A. Fisher in the 1920s and 1930s to note the importance of the physical act of randomization in experiments. Fisher first did this in the context of experimental design in his 1925 book, expanded on the issue in a 1926 article for agricultural researchers, and developed it more fully and for a broader audience in his 1935 book *The Design of Experiments* [for more on Fisher's role in the advocacy of randomization see Armitage (2003), Hall (2007), Preece (1990)]. As Reid (1982, p. 45) notes of Neyman: "On one occasion, when someone perceived him as anticipating the English statistician R.A. Fisher in the use of randomization, he objected strenuously:

> 'I treated *theoretically* an unrestrictedly randomized agricultural experiment and the randomization was considered as a prerequisite to probabilistic treatment of the results. This is not the same as the recognition that without randomization an experiment has little value irrespective of the subsequent treatment. The latter point is due to Fisher, and I consider it as one of the most valuable of Fisher's achievements.'[3]

This gap between Neyman and Fisher points to the fact that there was something absent from the Neyman mathematical formulation in 1923, which was added later, even though the symbolic formulation was complete in 1923. What those symbols *meant* changed. And in these changes lies what is causal about the Neyman-Rubin model—i.e., a focus on the mechanism by which treatment is assigned.

The Neyman-Rubin model is more than just the math of the original Neyman model. Obviously, it does not rely on an urn-model motivation for the observed potential

[3] Also see Rubin (1990, p. 477).

outcomes, but instead, for experiments, a motivation based on the random assignment of treatment. And for observational studies, one relies on the assumption that the assignment of treatment can be treated as if it were random. In either case, the mechanism by which treatment is assigned is of central importance. And the realization that the primacy of the assignment mechanism holds true for observational data no less than for experimental is due to Rubin (1974). This insight has been turned into a motto: "No causation without manipulation" (Holland 1986).

Although the original article was written in Polish, Neyman's work was known in the English-speaking world (Reid 1982), and in 1938 Neyman moved from Poland to Berkeley. It is thus unsurprising that the Neyman model quickly became the standard way of describing potential outcomes of randomized experiments (e.g., Anscombe 1948; Kempthorne 1952, 1955; McCarthy 1939; Pitman 1937; Welch 1937). The most complete discussion I know of before Rubin's work is Scheffé (1956). And a simplified version of the model even appears in an introductory textbook in the 1960s (Hodges & Lehmann 1964, sec. 9.4).[4]

The basic setup of the Neyman model is very simple. Let Y_{i1} denote the potential outcome for unit i if the unit receives treatment, and let Y_{i0} denote the potential outcome for unit i in the control regime. The treatment effect for observation i is defined by $\tau_i = Y_{i1} - Y_{i0}$. Causal inference is a missing data problem because Y_{i1} and Y_{i0} are never both observed. This remains true regardless of the methodology used to make inferential progress—regardless of whether we use quantitative or qualitative methods of inference. The fact remains that we cannot observe both potential outcomes at the same time.

Some assumptions have to be made to make progress. The most compelling are offered by a randomized experiment. Let T_i be a treatment indicator: 1 when i is in the treatment regime and 0 otherwise. The observed outcome for observation i is then:

$$Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}. \qquad 1.$$

Note that in contrast to the usual regression assumptions, the potential outcomes, Y_{i0} and Y_{i1}, are fixed quantities and not random variables, and that Y_i is only random because of treatment assignment.

Extensions to the case of multiple discrete treatment are straightforward (e.g., Imbens 2000; Rosenbaum 2002, pp. 300–2). Extensions to the continuous case are possible but lose the nonparametric nature of the Neyman model (see Imai & van Dyk 2004).

Experimental Data

In principle, if assignment to treatment is randomized, causal inference is straightforward because the two groups are drawn from the same population by construction, and treatment assignment is independent of all baseline variables. The distributions of both observed and unobserved variables between treatment and control groups are equal—i.e., the distributions are balanced. This occurs with arbitrarily high probability as the sample size grows large.

Treatment assignment is independent of Y_0 and Y_1—i.e., $\{Y_{i0}, Y_{i1} \perp\!\!\!\perp T_i\}$, where $\perp\!\!\!\perp$ denotes independence. In other words, the distributions of both of the potential outcomes (Y_0, Y_1) are the same for treated ($T = 1$) and control ($T = 0$). Hence, for $j = 0, 1$,

$$E(Y_{ij}|T_i = 1) = E(Y_{ij}|T_i = 0), \qquad 2.$$

where the expectation is taken over the distribution of treatment assignments. This equation states that the distributions of potential outcomes in treatment and control groups are the same in expectation. But for treatment observations one observes T_{i1} and for control observations T_{i0}. Treatment status filters which of the two potential outcomes we observe (Equation 1) but does not change them.

[4]The philosopher David Lewis (1973) is often cited for hypothetical counterfactuals and causality, and it is sometimes noted that he predated, by a year, Rubin (1974). The Neyman model predates Lewis.

ATE: average treatment effect

The average treatment effect (ATE) is defined to be:

$$\tau = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0)$$
$$= E(Y_i|T_i = 1) - E(Y_i|T_i = 0). \quad 3.$$

Equation 3 can be estimated consistently by simply taking the difference between two sample means because randomization ensures that the potential outcomes in treatment and control groups have the same distributions in expectation. This implies that randomization ensures that assignment to treatment will not be associated with any potentially confounding variable—i.e., with any pretreatment variable associated with the outcome.

One of the assumptions which randomization by itself does not justify is that "the observation on one unit should be unaffected by the particular assignment of treatments to the other units" (Cox 1958, sec. 2.4). "No interference between units" is often called the stable unit treatment value assumption (SUTVA). SUTVA implies that the potential outcomes for a given unit do not vary with the treatments assigned to any other unit, and that there are not different versions of treatment (Rubin 1978). SUTVA is a complicated assumption that is all too often ignored.

Brady (2008) describes a randomized welfare experiment in California where SUTVA is violated. In the experiment, teenage girls in the treatment group had their welfare checks reduced if they failed to obtain passing grades in school. Girls in the control group did not face the risk of reduced payments. However, some girls in the control group thought that they were in the treatment group, probably because they knew girls in that group (Mauldon et al. 2000). Therefore, the experiment probably underestimated the effect of the treatment.

Some researchers erroneously think SUTVA is another term for the assumption usually made in regression models that the disturbances of different observations are independent of one another. A hint of the problem can be seen by noting that ordinary least squares (OLS) is still unbiased under the usual assumptions even if multiple draws from the disturbance are not independent of each other. When SUTVA is violated, however, an experiment will not generally yield unbiased estimates (Cox 1958). In the usual regression setup, the correct specification assumption deals with SUTVA violations: It is implicitly assumed that if there are SUTVA violations, we have the correct model for them so that conditional independence holds—i.e., $E(\epsilon|X) = 0$, where ϵ is the regression disturbance and X represents the observed variables.

Even with randomization, the usual OLS regression assumptions are not satisfied. Indeed, without further assumptions, the multiple regression estimator is biased. Asymptotically the bias vanishes in some cases but need not with cluster randomized experiments (Middleton 2008). The regression standard errors can be severely biased, and the multiple regression estimator may have higher asymptotic variance than simply estimating Equation 3 (for details see Freedman 2008b,c). Intuitively, the problem is that generally, even with randomization, the treatment indicator and the disturbance will be strongly correlated. Randomization does not imply, as OLS assumes, a linear additive treatment effect where the coefficients are constant across units. Random effects do not solve the problem. Linear additivity remains, and the heterogeneity of the causal effect must be modeled. But the model may be wrong. For example, the effect may not vary normally as is commonly assumed, and it may be strongly related to other variables in the model. Researchers should be extremely cautious about using multiple regression to adjust experimental data. Unfortunately, there is a tendency to use it freely. This is yet another sign, as if one more were needed, of how ingrained the regression model is in our quantitative practice.

Unlike multiple regression, random assignment of treatment is sufficient for simple bivariate regression to be an unbiased estimator for Equation 3. The simple regression estimator is obtained by running a regression of the observed response Y on the assignment variable T with an intercept. The standard errors of this

estimator are, however, generally incorrect because the standard regression formulas assume homoscedasticity. Alternative variance estimators that adjust for heteroscedasticity may be used. An obvious alternative is to use the variance estimator $\frac{\hat{v}_t}{n_t} + \frac{\hat{v}_c}{n_c}$, where $\hat{v}_t$ is the sample variance for the treatment observations, n_t is the number of treatment observations, and the subscript c denotes analogous quantities for the control group.

The only stochastic thing in the Neyman-Rubin framework is the assignment to treatment. The potential outcomes are fixed. This is exactly the opposite of many econometric treatments, where all of the regressors (including the treatment indicator) are considered to be fixed, and the response variable Y is considered to be a random variable with a given distribution. None of that is implied by randomization, and indeed randomization explicitly contradicts it because one of the regressors (the treatment indicator) is explicitly random. Adding to the confusion is the tendency of some texts to refer to the fixed-regressors design as an experiment when that cannot possibly be the case.

In many modern treatments of OLS, X is stochastic, but that raises additional questions. Except for the randomly assigned treatment indicator, what makes the X covariates random? And if the data are a random sample (so, clearly, X is random), then there are two distinct sources of randomness: (*a*) treatment assignment; (*b*) sampling from a population. These are distinct entities, and one could be interested in either sample or population estimates—e.g., sample average treatment effects (SATE) or population average treatment effects (PATE). Sample estimates ignore the second source of randomness, and the population estimates take both into account. In the case of random sampling, SATE generally has less variance than PATE but certainly no more (Imbens 2004). Without assumptions in addition to random assignment and random sampling, one is not led to the usual regression variance formulas.

A parallel argument holds if one wants to consider the potential outcomes to be random and not fixed. What are the source and model of this randomness? Without additional information, it is most natural to consider that the potential outcomes are fixed because in a randomized experiment the only aspect that we *know* is random is treatment assignment. In the case of random potential outcomes, one can always conduct an analysis conditional on the data at hand, such as SATE, which ignores the second source of randomness. Of course, the conditional inference (e.g., SATE) may lead to a different inference than the unconditional inference. Without assumptions (such as random sampling), the sample contains no information about the PATE beyond the SATE. Note that if the potential outcomes are random, but we condition on the observed potential outcomes and so treat them as fixed, questions about the role of conditioning and inference arise, which go back to Neyman and Fisher. If the random error is independent of treatment assignment, this situation is analogous to the case of a 2×2 table where one margin is fixed and we analyze the data as if both margins are fixed (Lehmann 1993; Rosenbaum 2005, sec. 2.5–2.9).

ITT: intention to treat

Even in an experimental setup, much can go wrong that requires statistical adjustment (e.g., Barnard et al. 2003). A common problem is compliance. For example, a person assigned to treatment may refuse it. This person is said to have crossed over from treatment to control. A person assigned to control may find some way to receive treatment nevertheless, which is another form of crossover.

When there are compliance issues, Equation 3 defines the intention-to-treat (ITT) estimand. Although the concept of ITT dates earlier, the phrase probably first appeared in print in 1961 (Hill 1961, p. 259). Moving beyond the ITT to estimate the effect of treatment on the units that actually received it can be difficult. ITT measures the effect of assignment rather than treatment itself, and estimates of ITT are unbiased even with crossover. The obvious benefit is that ITT avoids bias by taking advantage of the experimental design.

The simplest compliance problem is one in which every unit assigned to control accepts

ETT: effect of treatment on the treated

control, but some units assigned to treatment decline treatment and follow the control protocol instead. This is called single crossover. In this case, the Neyman-Rubin model can easily handle the issue. Progress is made by assuming that there are two types of units: compliers and never-treat. A complier follows her assignment to either treatment or control. Compliers have two potential outcomes, which are observed as in Equation 1. However, a never-treat unit is assumed to have only one response, and this response is observed regardless of whether the unit is randomized to receive treatment or control.

With this simple model in place, we have five different parameters:

- the proportion of compliers in the experimental population (α)
- the average response of compliers assigned to treatment ($\bar{W}$)
- the average response of compliers assigned to control ($\bar{C}$)
- the difference between $\bar{W}$ and $\bar{C}$, which is the average effect of treatment on the compliers ($\bar{R}$)
- the average response of never-treat units assigned to control ($\bar{Z}$)

All five of these parameters can be estimated. α can be estimated by calculating the proportion of compliers observed in the treatment group. Because of randomization, this proportion is an unbiased estimate of the proportion of compliers in control as well. The average response of compliers to treatment, $\bar{W}$, is simply the average response of compliers in the treatment group. And $\bar{Z}$, the average response of never-treat units to control, is estimated by the average response among units in the treatment group who refused treatment.

This leaves $\bar{C}$ and $\bar{R}$. For $\bar{R}$, note that the control group contains a mix of compliers and never-treat units. We do not know the type of any given unit in control, but we know (in expectation) the *proportion* of each there must be in control because we can estimate this proportion in the treated group.

Recall that α denotes the proportion of compliers in the experimental population, and assume $\alpha > 0$. Under the model, the proportion of never-treat units must be $1 - \alpha$. Denote the average observed responses in treatment and control by $\bar{Y}^t$, $\bar{Y}^c$; these are sample quantities that are directly observed. Since the treatment and control groups are exchangeable because of random assignment,

$$E(\bar{Y}^c) = \alpha\bar{C} + (1-\alpha)\bar{Z}.$$

Therefore,

$$\bar{C} = \frac{E(\bar{Y}^c) - (1-\alpha)\bar{Z}}{\alpha}.$$

An obvious estimator for $\bar{C}$ is

$$\hat{\bar{C}} = \frac{\bar{Y}^c - (1-\hat{\alpha})\hat{\bar{Z}}}{\hat{\alpha}}.$$

Then the only remaining quantity is $\bar{R}$, the average effect of treatment on the compliers—i.e., the effect of treatment on the treated (ETT). This can be estimated by

$$\hat{\bar{W}} - \hat{\bar{C}} = \frac{\bar{Y}^t - \bar{Y}^c}{\hat{\alpha}}. \qquad 4.$$

Note how simple and intuitive Equation 4 is. The estimated average effect of treatment on the treated is calculated by dividing the ITT estimator by the compliance rate. Because this rate is less than or equal to 1 and, by assumption, above 0, ETT will be greater than or equal to ITT, and both will have the same sign.

Equation 4 is the same as two-stage least squares where the instrument is the random assignment to treatment. The canonical citation for this estimator is Angrist et al. (1996); they provide a more general derivation. The discussion above implicitly satisfies the assumptions they outline. For other discussions see Angrist & Imbens (1994), Bloom (1984), Freedman (2006), and Sommer & Zeger (1991).

When the compliance problem has a more complicated structure (e.g., when there is two-way crossover), it is difficult to make progress without making strong structural assumptions (Freedman 2006). We return to the issue of compliance in a later section, when we discuss the get-out-the-vote controversy.

Observational Data

In an observational setting, unless something special is done, treatment and nontreatment groups are almost never balanced because the two groups are not ordinarily drawn from the same population. Thus, a common quantity of interest is the average treatment effect for the treated (ATT):

$$\tau|(T=1) = E(Y_{i1}|T_i=1) - E(Y_{i0}|T_i=1), \quad 5.$$

where the expectation is taken over the distribution of treatment assignments. Equation 5 cannot be directly estimated because Y_{i0} is not observed for the treated. Progress can be made by assuming that selection for treatment depends on observable covariates denoted by X. Then, one can assume that conditional on X, treatment assignment is unconfounded. In other words, the conditional distributions of the potential outcomes are the same for treated and control: $\{Y_0, Y_1 \perp\!\!\!\perp T\}|X$.

Following Rosenbaum & Rubin (1983), we say that treatment assignment is strongly ignorable given a vector of covariates X if unconfoundedness and common overlap hold:

$$\{Y_0, Y_1 \perp\!\!\!\perp T\}|X$$
$$0 < \Pr(T=1|X) < 1$$

for all X. Heckman et al. (1998) show that for ATT, the unconfoundedness assumption can be weakened to conditional mean independence between the potential outcomes Y_{ij} and the treatment indicator T_i given X_i (also see Abadie & Imbens 2006a).

The common overlap assumption ensures that some observed value of X does not deterministically result in a given observation being assigned to treatment or control. If such deterministic treatment assignments were to occur, it would not be possible to identify the treatment effect. For example, if women were never treated and men always treated, it would not be possible to obtain an unbiased estimate of the average treatment effect (ATE) without an additional assumption.[5]

Given strong ignorability, following Rubin (1974, 1977) we obtain

$$E(Y_{ij}|X_i, T_i=1) = E(Y_{ij}|X_i, T_i=0). \quad 6.$$

ATT: average treatment effect for the treated

Equation 6 is the observational equivalent of Equation 2. Equation 6 is a formalization of the "as if random" assumption made in observational studies. Once some observable variables have been conditioned upon, analysis can continue as if treatment were randomly assigned. A key goal is to obtain results for observational data that were demonstrated to hold given random assignment in the previous section.

By conditioning on observed covariates, X_i, treatment and control groups are balanced—i.e., the distributions of the potential outcomes between treatment and control groups are the same. When it comes to potential outcomes, the only difference between the two groups is the potential outcomes we observe, Y_i or Y_0. The ATE for the treated is estimated as

$$\tau|(T=1) = E\{E(Y_i|X_i, T_i=1) - E(Y_i|X_i, T_i=0)|T_i=1\}, \quad 7.$$

where the outer expectation is taken over the distribution of $X_i|(T_i=1)$, which is the distribution of X in the treated group.

Note that the ATT estimator is changing how individual observations are weighted, and that observations outside of common support receive zero weights. That is, if some covariate values are only observed for control observations, those observations will be irrelevant for estimating ATT and are effectively dropped. Therefore, the overlap assumption for ATT only requires that the support of X for the treated observations be a subset of the support of X for control observations. More generally, one would also want to drop treatment observations if they have covariate values that do not overlap with control observations (Crump et al. 2006). In such cases, it is unclear exactly what estimand one is estimating because it is no longer ATT, as some treatment observations

[5] We could assume that sex is independent of the potential outcomes. Women in the control group could then be valid counterfactuals for men in treatment given the Y of interest. Such additional exclusion assumptions are not required if strong ignorability holds.

have been dropped along with some control observations.

It is often jarring for people to hear that observations are being dropped because of a lack of covariate overlap. Our intuition against dropping observations comes from what happens with experimental data, where homogeneity between treatment and control is guaranteed by randomization so a larger sample is obviously better than a smaller one. But with observational data, dropping observations that are outside of common support not only reduces bias but can also reduce the variance of our estimates. This may be counterintuitive, but note that our variance estimates are a function of both sample size and unit heterogeneity—e.g., in the regression case, of the sample variance of X and the mean square error. Dropping observations outside of common support and conditioning as in Equation 7 helps to improve unit homogeneity and may actually reduce our variance estimates (Rosenbaum 2005). Rosenbaum 2005 also shows that, with observational data, minimizing unit heterogeneity reduces both sampling variability and sensitivity to unobserved bias. With less unit heterogeneity, larger unobserved biases need to exist to explain away a given effect. And although increasing the sample size reduces sampling variability, it does little to reduce concerns about unobserved bias. Thus, maximizing unit homogeneity to the extent possible is an important task for observational methods.[6]

The key assumption being made here is strong ignorability. Even thinking about this assumption presupposes some rigor in the research design. For example, is it clear what is pre- and what is posttreatment? If not, one cannot even form the relevant questions. The most useful of those questions may be the one suggested by Dorn (1953, p. 680), who proposed that the designer of every observational study should ask, "How would the study be conducted if it were possible to do it by controlled experimentation?" This clear question also appears in Cochran's (1965) famous Royal Statistical Society discussion paper on the planning of observational studies of human populations. Researchers in the tradition of the Neyman-Rubin model routinely ask Dorn's question of themselves and their students. The question forces the researcher to focus on a clear manipulation and then on the selection problem at hand. Only then can one even begin to think clearly about how plausible the strong ignorability assumption may or may not be. Because most researchers do not propose an answer to this question, it is difficult to think clearly about the underlying assumptions being made in most applications in the social sciences because it is not clear what the researcher is trying to estimate.

For the moment, let us assume that the researcher has a clear treatment of interest and a set of confounders that may reasonably ensure conditional independence of treatment assignment. At that point, one needs to condition on these confounders (denoted by X). But we must remember that selection on observables is a large concession, which should not be made lightly. It is of far greater relevance than the following technical discussion on the best way to condition on covariates.

In other words, the identification assumption for both OLS and matching is the same: selection on observables. Both also rely on the stable unit treatment value assumption (SUTVA) and have similar restrictions on the use of posttreatment variables. Despite their differences, they have more in common than most applied researchers in political science realize. Therefore, the identification assumption—e.g., selection on observables—should receive more attention than is often the case in the literature. Authors, even when they have natural experiments, spend insufficient effort justifying this assumption [for a review and evaluation of a number of natural experiments and their "as

[6]There is a trade-off between having a smaller number of more homogeneous observations and a larger number of more heterogeneous observations. Whether dropping a given observation actually increases the precision of the estimate depends on how different this observation is from the observations that remain and how sensitive the estimator is to heterogeneity (see Rosenbaum 2005 for formal details).

if random" assumptions, see Dunning (2008)]. Obviously, matching is nonparametric whereas OLS is not. This is an important distinction because asymptotically matching does not make a functional form assumption in addition to the selection-of-observables assumption (Abadie & Imbens 2006a). OLS, however, does make additional assumptions; it assumes linear additivity.

MATCHING METHODS

The most straightforward and nonparametric way to condition on X is to exactly match on the covariates. This is an old approach, dating back at least to Fechner (1966 [1860]), the father of psychophysics. This approach is often impossible to implement in finite samples if the dimensionality of X is large—i.e., exact matches are not found in a given sample. And exact matching is not possible to implement even asymptotically if X contains continuous covariates. Thus, in general, alternative methods must be used.

Various forms of matching have been used for some time, for example (Chapin 1938, Cochran 1953, Greenwood 1945). Two common approaches today are propensity score matching (Rosenbaum & Rubin 1983) and multivariate matching based on Mahalanobis distance (Cochran & Rubin 1973; Rubin 1979, 1980).

Mahalanobis and Propensity Score Matching

The most common method of multivariate matching is based on Mahalanobis distance (Cochran & Rubin 1973; Rubin 1979, 1980). The Mahalanobis distance between any two column vectors is

$$md(X_i, X_j) = \{(X_i - X_j)' S^{-1} (X_i - X_j)\}^{\frac{1}{2}},$$

where S is the sample covariance matrix of X. To estimate ATT, one matches each treated unit with the M closest control units, as defined by this distance measure, $md()$. Matching with replacement results in the estimator with the lowest conditional bias (Abadie & Imbens 2006a). [Alternatively, one can use optimal full matching (Hansen 2004, Rosenbaum 1991), which may have lower variance. But this decision is separate from the choice of a distance metric.] If X consists of more than one continuous variable, multivariate matching estimates contain a bias term that does not asymptotically go to zero at $\sqrt{n}$ (Abadie & Imbens 2006a).

EPBR: equal percent bias reduction

An alternative way to condition on X is to match on the probability of assignment to treatment, known as the propensity score.[7] As one's sample size grows large, matching on the propensity score produces balance on the vector of covariates X (Rosenbaum & Rubin 1983).

Given strong ignorability, Rosenbaum & Rubin (1983) prove

$$\tau|(T=1) = E\{E(Y_i|e(X_i), T_i = 1) - E(Y_i|e(X_i), T_i = 0)|T_i = 1\},$$

where the outer expectation is taken over the distribution of $e(X_i)|(T_i = 1)$. Under these assumptions, the propensity score can be used to provide an unbiased estimate of ATE as well.

Propensity score matching usually involves matching each treated unit to the nearest control unit on the unidimensional metric of the propensity score vector. [Optimal matching might sometimes match treated units to nonnearest control units in order to minimize the overall distance (Hansen 2004, Rosenbaum 1991).] Because the propensity score is generally unknown, it must be estimated. If the propensity score is estimated by logistic regression, as is typically the case, much is to be gained by matching not on the predicted probabilities (bounded between zero and one) but on the linear predictor: $\hat{\mu} = X\hat{\beta}$. Matching on the linear predictor avoids compression of propensity scores near zero and one (Rosenbaum & Rubin 1985). Moreover, the linear predictor is often more nearly normally distributed, which is of some importance given the "equal percent bias reduction" (EPBR) theoretical results discussed below.

[7] The first estimator of treatment effects to be based on a weighted function of the probability of treatment was the Horvitz-Thompson statistic (Horvitz & Thompson 1952).

EQUAL PERCENT BIAS REDUCTION

Affinely invariant matching methods, such as Mahalanobis metric matching and propensity score matching (if the propensity score is estimated by logistic regression), are equal percent bias reducing if all of the covariates used have ellipsoidal distributions (Rubin & Thomas 1992)—e.g., distributions such as the normal or *t*—or if the covariates are discriminant mixtures of proportional ellipsoidally symmetric (DMPES) distributions (Rubin & Stuart 2006). Note that DMPES defines a limited set of mixtures—in particular, countably infinite mixtures of ellipsoidal distributions where (*a*) all inner products are proportional and (*b*) the center of each constituent ellipsoidal distribution is such that all best linear discriminants between any two components are also proportional.

To formally define EPBR, let Z be the expected value of X in the matched control group. Then, as outlined by Rubin (1976a), a matching procedure is EPBR if

$$E(X\,|\,T=1) - Z = \gamma\{E(X\,|\,T=1) - E(X\,|\,T=0)\}$$

for a scalar $0 \leq \gamma \leq 1$. In other words, a matching method is EPBR for X when the percent reduction in the biases of each of the matching variables is the same. One obtains the same percent reduction in bias for any linear function of X if and only if the matching method is EPBR for X. Moreover, if a matching method is not EPBR for X, the bias for some linear function of X is increased even if all univariate covariate means are closer in the matched data than in the unmatched (Rubin 1976a).

Mahalanobis distance and propensity score matching can be combined in various ways (Rubin 2001). Rosenbaum & Rubin (1985) show that, in finite samples, it is useful to combine the two matching methods because doing so reduces covariate imbalance and mean squared error in the causal estimate more effectively than using either method alone. The improvements occur because the propensity score is a balancing score only asymptotically. In finite samples, some covariate imbalances will remain, which another matching method can help adjust.

DMPES distributions: discriminant mixtures of proportional ellipsoidally symmetric distributions

Matching methods based on the propensity score (estimated by logistic regression), Mahalanobis distance, or a combination of the two have appealing theoretical properties if covariates have ellipsoidal distributions—e.g., distributions such as the normal or *t*. If the covariates are so distributed, these methods (more generally, affinely invariant matching methods[8]) have the property of EPBR (Rubin 1976a,b; Rubin & Thomas 1992).[9] This property, formally defined in the sidebar "Equal Percent Bias Reduction," ensures that matching methods will reduce bias in all linear combinations of the covariates. If a matching method is not EPBR, then that method will, in general, increase the bias for some linear function of the covariates even if all univariate means are closer in the matched data than the unmatched (Rubin 1976a).

A significant shortcoming of these common matching methods is that they may (and in practice, frequently do) make balance worse across measured potential confounders. These methods may make balance worse even if the distribution of covariates is ellipsoidally symmetric, because EPBR is a property that holds in expectation. That is, even if the covariates have elliptic distributions, finite samples may not conform to ellipticity, and hence Mahalanobis distance may not be optimal because the matrix used to scale the distances, the sample covariance matrix of X, may not be sufficient to account for all of the differences between the distributions of the covariates in X. In finite samples, there may be more differences between the distributions of covariates than just means and variances—e.g., the other moments may differ as well. [On Mahalanobis distance and distributional considerations, see Mitchell & Krzanowski (1985, 1989).] Moreover, if covariates are neither ellipsoidally symmetric nor discriminant mixtures of proportional ellipsoidally symmetric (DMPES) distributions, propensity score matching has good theoretical

[8]Affine invariance means that the matching output is invariant to matching on X or an affine transformation of X.

[9]The EPBR results of Rubin & Thomas (1992) have been extended by Rubin & Stuart (2006) to the case of discriminant mixtures of proportional ellipsoidally symmetric distributions. This extension is important, but it is restricted to a limited set of mixtures.

properties only if the true propensity score model is known with certainty and the sample size is large.

The EPBR property itself is limited and not always desirable. Consider a substantive problem in which it is known, based on theory, that one covariate has a large nonlinear relationship with the outcome while another does not—e.g., $Y = X_1^4 + X_2$, where $X > 1$ and where both X_1 and X_2 have the same distribution. In such a case, covariate imbalance in X_1 will be generally more important than X_2 because the response surface (i.e., the model of Y) is more sensitive to changes in X_1 than X_2.

Genetic Matching

Given these limitations, it may be desirable to use a matching method that algorithmically imposes certain properties when the EPBR property does not hold. One method that does this while keeping the estimand constant is genetic matching (GenMatch) (Diamond & Sekhon 2005, Sekhon 2009). GenMatch automatically finds the set of matches that minimizes the discrepancy between the distribution of potential confounders in the treated and control groups. That is, covariate balance is maximized. GenMatch is a generalization of propensity score and Mahalanobis distance matching. It has been used by a variety of researchers (e.g., Bonney et al. 2007, Boyd et al. 2008, Eggers & Hainmueller 2008, Gilligan & Sergenti 2008, Gordon & Huber 2007, Heinrich 2007, Herron & Wand 2007, Korkeamäki & Uuistalo 2009, Lenz & Ladd 2006, Raessler & Rubin 2005, Woo et al. 2008). The method uses a genetic algorithm (Mebane & Sekhon 2009, Sekhon & Mebane 1998) to optimize balance as much as possible given the data. GenMatch is nonparametric and does not depend on knowing or estimating the propensity score, but the method is improved when a propensity score is incorporated. Diamond & Sekhon (2005) use this algorithm to show that the long-running debate between Dehejia & Wahba (1997, 1999, 2002; Dehejia 2005) and Smith & Todd (2005a,b, 2001) is largely a result of the use of models that do not produce good balance—even if some of the models get close, by chance, to the experimental benchmark of interest. They show that GenMatch is able to quickly find good balance and to reliably recover the experimental benchmark. Sekhon & Grieve (2008) show that for a clinical intervention of interest in the matching literature, pulmonary artery catheterization, applying GenMatch to an observational study replicates the substantive results of a corresponding randomized controlled trial, unlike the extant literature.

A difficult question all matching methods must confront is how to measure covariate balance. Users of propensity score matching iterate between tweaking the specification of their propensity score model and then checking the covariate balance. Researchers stop when they are satisfied with the covariate balance they have obtained or when they tire. One process for cycling between checking for balance on the covariates and reformulating the propensity score model is outlined by Rosenbaum & Rubin (1984). GenMatch is an alternative to this process of reformulating the propensity score model, and like other forms of matching, it is agnostic about how covariate balance is measured because this is an open research question. Therefore, the GenMatch software (Sekhon 2009) offers a variety of ways to measure covariate balance, many of which rely on cumulative probability distribution functions. By default, these statistics include paired t-tests, and univariate and multivariate Kolmogorov-Smirnov tests. Various descriptive statistics based on empirical-QQ plots are also offered. The statistics are not used to conduct formal hypothesis tests, because no measure of balance is a monotonic function of bias in the estimand of interest and because we wish to maximize balance without limit (Imai et al. 2008, Sekhon 2006). GenMatch can maximize balance based on a variety of predefined measures of balance or any measure the researcher may wish to use, such as the Kullback-Leibler divergence measure, which is popular in information theory and image processing (Kullback & Leibler 1951). For details see Sekhon (2009).

GET-OUT-THE-VOTE CONTROVERSY

In a landmark study of various get-out-the-vote (GOTV) interventions, Gerber & Green (2000) reported results from a field experiment they conducted in New Haven in 1998. Revisiting Eldersveld (1956), Gerber & Green 2000 examined the relative effectiveness of various GOTV appeals, including short nonpartisan telephone calls, direct mail, and personal canvassing. They found that "[v]oter turnout was substantially increased by personal canvassing, slightly by direct mail, and not at all by telephone calls" (Gerber & Green 2000, p. 653). These results held for both ITT (intention to treat) and ETT (effect of treatment on the treated). The noncompliance problem in this experiment consists of only single crossover—i.e., there are two types of units, compliers and never-treat. With random assignment of ITT, ETT can be estimated consistently with the two-stage least squares approach of Equation 4, which Gerber & Green used.

Imai (2005) argues that the attempt to randomly assign treatment in the Gerber & Green study was not successful, and hence, the field experiment should be analyzed using observational methods alone. It is argued that neither ITT nor ETT could be estimated without adjustment. Imai uses propensity score matching to estimate ETT. Imai assumes that once a set of observables has been matched upon using his propensity score, the outcomes of compliers assigned to treatment can be compared with the outcomes of units assigned to control to estimate ETT. The inferential problem is that the control group consists of both never-treats and compliers, whereas the units assigned to treatment who received treatment are all compliers.

The observables used by Imai were drawn from the usual voter registration files. There were six covariates for each subject. The indicator variables were as follows: turnout in the prior election, 1996; new voter registrant; major party registrant; and single-voter household. The two additional covariates were the age of the subject and the ward of residence.

Imai argues that contrary to the original findings, short nonpartisan telephone appeals did have a significant positive effect on turnout. Green and Gerber responded in various articles (Arceneaux et al. 2006, Gerber & Green 2005), and Bowers and Hansen entered the debate using alternative methods (Bowers & Hansen 2005, Hansen & Bowers 2009) that reconfirmed the substantive findings of Gerber & Green (2000).

Imai performed an invaluable service by prompting Gerber and Green to find and correct a number of data-processing errors in the original Gerber & Green (2000) study.[10] Imai also performed an important service by pointing out that at the level of individuals, the experiment did not appear to be randomized successfully even after data-processing errors were corrected—i.e., covariate imbalances between treatment and control were greater than one would expect by chance. In the original study, the data were analyzed as if individuals were randomized even though randomization was actually by household. Prompted by Imai, subsequent randomization checks were performed at the household level once household identifiers were released.

Consistent with the findings of Gerber & Green (2000), all analysts aside from Imai have concluded that short nonpartisan telephone calls are not effective. This holds in the original data for the New Haven study (Bowers & Hansen 2005, Gerber & Green 2005), the corrected data (Gerber & Green 2005, Hansen & Bowers 2009), and subsequent large-scale field experiments conducted in Michigan and Iowa (Arceneaux et al. 2006).

This exchange highlights an important lesson: When analyzing any experiment, one

[10]According to Gerber & Green (2005), there were data-processing errors related to: (*a*) imperfect matches between names on the original master file and the names returned by canvassers; (*b*) a failure of communication with the phone bank about which treatment groups were to be assigned the GOTV appeal; (*c*) data manipulation errors that resulted in some subjects in the control group being incorrectly recorded as treatment subjects.

should stay as close to the experimental design as possible. This holds even if one conjectures that randomization has not fully balanced the covariates in the given sample. Discarding the experimental design and reverting to purely observational methods fails to result in unbiased estimates of the effectiveness of short nonpartisan telephone calls.

Because treatment in the original New Haven experiment was actually randomized at the level of households and not individuals, all randomization checks should be conducted at the household level. Failing to do so results in a spurious finding that randomization had not balanced the observable covariates, when in fact it had. And, ideally, variance estimates should take into account that randomization was done at the household level, although in this example this does not appear to make a significant substantive difference because the number of households is large.

With the corrected data, when the randomization checks are performed at the level of household, one finds that randomization was successful (Gerber & Green 2005, Hansen & Bowers 2009). Therefore, no method is needed to correct for any randomization issues. Before the household data were available and before it was known by Imai or Bowers & Hansen that randomization was done by household, it was found that if matching was used to simply strengthen the randomization—i.e., the randomization was not ignored—the original Gerber & Green results were recovered (Bowers & Hansen 2005). The simplest method of strengthening the randomization is to use stratification: to apply the estimator in Equation 4 within strata defined by observed confounders. Within each stratum, the confounders used to define the strata obviously cannot be an issue (if the covariates are homogeneous within strata).

Even if the original New Haven dataset is examined, and randomization is ignored, Imai's results are not robust to slight changes in methodology such as correcting his biased variance estimates. Unconventionally, Imai reported not the full sample point estimate, but the average estimate from 500 bootstrap estimates. However, using the full sample point estimate results in a p-value that is not significant at conventional test levels, even if one uses Imai's bootstrap variance estimate (Gerber & Green 2005). But bootstrapping yields biased variance estimates for matching estimators (Abadie & Imbens 2006b). If one does not use the bootstrap but, for example, the Abadie & Imbens (2006a) approach to estimate the point and variance estimates, one does not obtain a significant estimate at conventional levels (the point estimate is 5.6%, and the Abadie-Imbens standard error is 3.2). The same holds if one uses Imai's own code but simply does one-to-one matching with replacement (Gerber & Green 2005).

Matching in this example fails at least two different placebo tests. Placebo tests are underused as robustness checks in observational studies. Such tests are the observational equivalent of giving a sugar pill to a patient in the control group in a clinical trial. We know a priori that such a pill should have a zero treatment effect because of our knowledge of the biochemical properties of sugar pills. Therefore, the biochemical effectiveness of the treatment of interest can be estimated by comparing it to the results from the placebo group. (Even if the placebo does have an effect, we know it cannot be because of any biochemical property of the pill itself, so the placebo group still serves as a useful benchmark against which to measure the treatment of interest.) In an observational placebo test, one attempts to find a stratum of data and an outcome for which the treatment effect is known with similar certainty. Then one tests to see if the observational method one is using is able to recover the result that is known a priori. In this fashion, one simultaneously checks both the selection-on-observables assumption and the estimator. In the present case there are two obvious placebo tests.

The first, which is the clearer one because it follows directly from the assumptions of the

matching estimator, is to estimate the causal effect of being assigned to treatment but never receiving it. Since being assigned to receive a telephone GOTV appeal but never receiving the appeal cannot logically have an effect on turnout, we have a clear placebo: the causal effect must logically be zero. The outcomes of never-treat units who were assigned to treatment are being compared with the outcomes of the never-treat units who were assigned to control. The control group, however, consists of units who would be never-treat if they were assigned to treatment and units who would be compliers. For a valid comparison, one has to find the never-treat in the control group to compare with the never-treat who are assigned to treatment. Imai's observational approach purports to solve this inferential problem, since he has to find the compliers in control to compare with the compliers in treatment. Unfortunately, the estimate produced for this placebo test by one-to-five propensity score matching, the type used by Imai (2005), is −5.6% with a standard error of 2.3 (Gerber & Green 2005).

A second placebo test is offered by considering whether telephone calls have a zero effect on *past* turnout. In this setup, one obviously does not match on previous turnout since that becomes the "outcome" of interest, but one does match on the turnout before the placebo outcome. This placebo test is most appropriate for the Michigan and Iowa experiments described by Arceneaux et al. (2006) because of the availability of turnout history during the past two elections. In these experiments, exact matching estimates ETT to be 1.61% with an Abadie-Imbens (Abadie & Imbens 2006a) standard error of 0.258.[11] Exact matching was used to condition on turnout in the election before, age, gender, competitiveness, and household size. As in the previous placebo test, matching claims to find an effect where none is logically possible.

[11]This was estimated using the Matching package (Sekhon 2009) for the R Project for Statistical Computing.

Both of these placebo tests, if conducted, would probably have given any analyst pause. But as is all too common, the selection-on-observables assumption is accepted readily—by reviewers, by readers, and most importantly by data analysts themselves. Placebo tests, even when they are possible as in the present case, are rarely conducted.

This behavior is consistent with what has been observed in other disciplines, including economics, epidemiology, and clinical medicine. Experimental results are rarely recovered by observational methods, placebo tests are usually not done, and when they are reported by some researcher to caution against the use of observational methods, such tests are usually ignored. This occurs even in cases where lives are at stake. Tens of thousands of women probably died because their physicians prescribed hormone replacement therapy based on observational studies (Freedman & Petitti 2005a,b).

The GOTV controversy is odd. And its oddity highlights our discipline's belief in models. In order to use a matching algorithm, one need not have discarded all information about the experiment and reverted to purely observational methods. The hybrid approach of Bowers & Hansen (2005) allows one to adjust for any imbalance that remains in the observed covariates while using the information in the randomization. Both this hybrid approach and two-stage least squares with covariates make the same identification assumption. Both assume that once we condition on X, we can proceed as if the treatment assigned in the experiment is random and as if the compliance model described in the previous section holds. The two methods just differ in how they condition on X: via a parametric model or via stratification or matching. In contrast, as stated before, matching alone makes the same identifying assumption as OLS. Both methods rely on the selection-on-observables identification assumption, and they differ in the extent to which they rely on functional form assumptions.

Given the results of this debate, it is clear that the selection-on-observables assumption is

not valid in this case. And there may be lessons of general interest:

1. ITT should always be reported, and going beyond ITT should be done only with care.
2. All data analysis should leverage the experimental design as much as possible.
3. Selection on observables and other identifying assumptions not guaranteed by the design should be considered incorrect unless compelling evidence to the contrary is provided.
4. Placebo tests should be conducted whenever possible, and observational studies without them should be marked down.

CONCLUSION

As a discipline, we value novelty. But we do not want to change radically. We like new twists that do not challenge our standard research practices. With both quantitative and qualitative methods, we hope that the next innovation will solve our inference problems. Since we have tried to mass produce science on the cheap, we should not be surprised that a tradition which relies on finding a valid design is not dominant.

These observations are not new. David Freedman has made similar comments over the years about our discipline in particular and the social sciences in general (e.g., Freedman 1995, 1999, 2008a). In one famous example, he contrasts our norms and methods with the case of John Snow and cholera, a prominent example of the success of observational methods for causal inference (Freedman 1991, 1999; Snow 1855; Vinten-Johansen et al. 2003). As early as the cholera outbreak of 1831–1832, the first to reach England, Snow doubted the miasma theory as it applied to cholera. In the outbreak of 1848, he decided to track the progress of the disease, and he was able to find the index case, John Harnold, and document its spread and natural history. In the 1850s, Snow accumulated data on the epidemics of 1853–1854 and analyzed the "grand experiment" that linked the disease to specific water suppliers. The Broad Street pump natural experiment occurred in 1854. In 1831, Snow had a hypothesis based on evidence, but no compelling design to make a rigorous causal inference. For a compelling set of natural experiments he had to wait for 1854. A young researcher today who waited that long to find the right design would soon be out of a job. Researchers know this and adapt.

It should be no surprise that the modeling enterprise is the dominant one. Unfortunately, as matching is gaining popularity, its ties to the Neyman-Rubin causal model and considerations of design are weakening. Rubin (2008) notes that "design trumps analysis," but designs for observational data cannot be mass produced. From hunger comes our belief in analysis by models, statistical or otherwise, matching or kernel estimation, maximum likelihood or Bayesian.

For most researchers, the math obscures the assumptions. Without an experiment, a natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive. This conclusion has implications for the kind of causal questions we are able to answer with some rigor. Clear, manipulable treatments and rigorous designs are essential. And the only designs I know of that can be mass produced with relative success rely on random assignment. Rigorous observational studies are important and needed. But I do not know how to mass produce them.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I thank Jake Bowers, David Freedman, Don Green, Ben Hansen, Shigeo Hirano, Walter Mebane, Jr., Donald Rubin, Jonathan Wand, and Rocío Titiunik for valuable comments and advice. I also thank an anonymous reviewer for extensive and extremely helpful comments. All errors are my responsibility.

LITERATURE CITED

Abadie A, Imbens GW. 2006a. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74:235–67

Abadie A, Imbens GW. 2006b. *On the failure of the bootstrap for matching estimators*. Work. Pap., Harvard Univ.

Angrist JD, Imbens GW. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62(2):467–75

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91(434):444–55

Anscombe FJ. 1948. The validity of comparative experiments. *J. R. Stat. Soc. Ser. A* 61:181–211

Arceneaux K, Gerber AS, Green DP. 2006. Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Polit. Anal.* 14(1):37–62

Armitage P. 2003. Fisher, Bradford Hill, and randomization. *Int. J. Epidemiol.* 32(6):925–28

Barnard J, Frangakis CE, Hill JL, Rubin DB. 2003. Principal stratification approach to broken randomized experiments: a case study of school choice vouchers in New York City. *J. Am. Stat. Assoc.* 98:299–323

Bloom HS. 1984. Accounting for no-shows in experimental evaluation designs. *Eval. Rev.* 8(2):225–46

Bonney J, Canes-Wrone B, Minozzi W. 2007. *Issue accountability and the mass public: the electoral consequences of legislative voting on crime policy*. Work. Pap., Dep. Polit., Princeton Univ.

Bowers J, Hansen B. 2005. *Attributing effects to a get-out-the-vote campaign using full matching and randomization inference*. **http://www.jakebowers.org/PAPERS/bowershansen03Apr05.pdf**

Boyd CL, Epstein L, Martin AD. 2008. *Untangling the causal effects of sex on judging*. Presented at Annu. Conf. Empirical Legal Stud., 2nd, New York, Nov. 9–10. Available at SSRN: **http://ssrn.com/abstract=1001748**

Brady H. 2008. Causation and explanation in social science. In *The Oxford Handbook of Political Methodology*, ed. JM Box-Steffensmeier, HE Brady, D Collier, pp. 217–70. New York: Oxford Univ. Press

Breiman L, Friedman J, Stone CJ, Olshen RA. 1984. *Classification and Regression Trees*. New York: Chapman & Hall

Chapin SF. 1938. Design for social experiments. *Am. Sociol. Rev.* 3(6):786–800

Chipman HA, George EI, McCulloch RE. 2006. *BART: Bayesian additive regression trees*. Work. Pap., Grad. School Business, Univ. Chicago

Christakis NA, Iwashyna TI. 2003. The health impact of health care on families: a matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses. *Soc. Sci. Med.* 57(3):465–75

Cochran WG. 1953. Matching in analytical studies. *Am. J. Public Health* 43:684–91

Cochran WG. 1965. The planning of observational studies of human populations (with discussion). *J. R. Stat. Soc. Ser. A* 128:234–55

Cochran WG, Rubin DB. 1973. Controlling bias in observational studies: a review. *Sankhya, Ser. A* 35:417–46

Cox DR. 1958. *Planning of Experiments*. New York: Wiley

Crump RK, Hotz VJ, Imbens GW, Mitnik OA. 2006. *Moving the goalposts: addressing limited overlap in estimation of average treatment effects by changing the estimand*. Work. Pap. Dep., Econ., Harvard Univ.

Dehejia R. 2005. Practical propensity score matching: a reply to Smith and Todd. *J. Econometrics* 125:355–64

Dehejia R, Wahba S. 1997. Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. In *Econometric methods for program evaluation*, R Dehejia, Ch. 1. PhD thesis, Harvard Univ.

Dehejia R, Wahba S. 1999. Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. *J. Am. Stat. Assoc.* 94(448):1053–62

Dehejia RH, Wahba S. 2002. Propensity score matching methods for nonexperimental causal studies. *Rev. Econ. Stat.* 84(1):151–61
Diamond A, Sekhon JS. 2005. *Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies*. Work. Pap., Dep. Econ., Harvard Univ.
Diprete TA, Engelhardt H. 2004. Estimating causal effects with matching methods in the presence and absence of bias cancellation. *Sociol. Methods Res.* 32(4):501–28
Dorn HF. 1953. Philosophy of inference from retrospective studies. *Am. J. Public Health* 43:692–99
Dunning T. 2008. Improving causal inference: strengths and limitations of natural experiments. *Polit. Sci. Q.* 61(2):282–93
Eggers A, Hainmueller J. 2008. *The value of political power: estimating returns to office in post-war British politics*. Work. Pap., Dep. Gov., Harvard Univ.
Eldersveld SJ. 1956. Experimental propaganda technique and voting behavior. *Am. Polit. Sci. Rev.* 50(1):154–65
Fechner GT. 1966 (1860). *Elements of Psychophysics*, Vol. 1. Transl. HE Adler, ed. DH Howes, EG Boring. New York: Rinehart & Winston. From German
Freedman DA. 1991. Statistical models and shoe leather. *Sociol. Methodol.* 21:291–313
Freedman DA. 1995. Some issues in the foundation of statistics. *Found. Sci.* 1:19–39
Freedman DA. 1999. From association to causation: some remarks on the history of statistics. *Stat. Sci.* 14:243–58
Freedman DA. 2004. On specifying graphical models for causation, and the identification problem. *Eval. Rev.* 26(4):267–93
Freedman DA. 2006. Statistical models for causation: What inferential leverage do they provide? *Eval. Rev.* 30:691–713
Freedman DA. 2008a. Oasis or mirage? *CHANCE Mag.* 21(1):59–61
Freedman DA. 2008b. On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.* 2(1):176–96
Freedman DA. 2008c. On regression adjustments to experimental data. *Adv. Appl. Math.* 40(2):180–93
Freedman DA. 2008d. Randomization does not justify logistic regression. *Stat. Sci.* 23(2):237–49
Freedman DA, Petitti DB. 2005a. Hormone replacement therapy does not save lives: comments on the Women's Health Initiative. *Biometrics* 61(4):918–20
Freedman DA, Petitti DB. 2005b. Invited commentary: How far can epidemiologists get with statistical adjustment? *Am. J. Epidemiol.* 162(5):415–18
Galiani S, Gertler P, Schargrodsky E. 2005. Water for life: the impact of the privatization of water services on child mortality. *J. Polit. Econ.* 113(1):83–120
Gerber AS, Green DP. 2000. The effects of canvassing, telephone calls, and direct mail on voter turnout: a field experiment. *Am. Polit. Sci. Rev.* 94(3):653–63
Gerber AS, Green DP. 2005. Correction to Gerber and Green (2000) replication of disputed findings, and reply to Imai (2005). *Am. Polit. Sci. Rev.* 99(2):301–13
Gilligan MJ, Sergenti EJ. 2008. Evaluating UN peacekeeping with matching to improve causal inference. *Q. J. Polit. Sci.* 3(2):89–122
Gordon S, Huber G. 2007. The effect of electoral competitiveness on incumbent behavior. *Q. J. Polit. Sci.* 2(2):107–38
Gosnell HF. 1927. *Getting Out the Vote: An Experiment in the Stimulation of Voting*. Chicago: Univ. Chicago Press
Gosnell HF. 1948. Mobilizing the electorate. *Ann. Am. Acad. Polit. Soc. Sci.* 259:98–103
Green D, Gerber A. 2002. Reclaiming the experimental tradition in political science. In *State of the Discipline*, Vol. III, ed. H Milner, I Katznelson, pp. 805–832. New York: W.W. Norton
Greenwood E. 1945. *Experimental Sociology: A Study in Method*. New York: King's Crown
Haavelmo T. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 1:1–12
Hall N. 2007. R. A. Fisher and his advocacy of randomization. *J. Hist. Biol.* 40(2):295–325
Hansen BB. 2004. Full matching in an observational study of coaching for the SAT. *J. Am. Stat. Assoc.* 99:609–18
Hansen BB, Bowers J. 2009. Attributing effects to a cluster randomized get-out-thevote campaign. *J. Am. Stat. Assoc.* In press

Heckman JJ, Ichimura H, Smith J, Todd P. 1998. Characterizing selection bias using experimental data. *Econometrica* 66(5):1017–98
Heinrich CJ. 2007. Demand and supply-side determinants of conditional cash transfer program effectiveness. *World Dev.* 35(1):121–43
Herron MC, Wand J. 2007. Assessing partisan bias in voting technology: the case of the 2004 New Hampshire recount. *Electoral Stud.* 26(2):247–61
Hill B. 1961. *Principles of Medical Statistics*. London: Lancet. 7th ed.
Hodges JL, Lehmann EL. 1964. *Basic Concepts of Probability and Statistics*. San Fransisco: Holden-Day
Holland PW. 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81(396):945–60
Horvitz DG, Thompson DJ. 1952. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47:663–85
Imai K. 2005. Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments. *Am. Polit. Sci. Rev.* 99(2):283–300
Imai K, King G, Stuart EA. 2008. Misunderstandings among experimentalists and observationalists about causal inference. *J. R. Stat. Soc. Ser. A* 171(2):481–502
Imai K, van Dyk DA. 2004. Causal inference with general treatment regimes: generalizing the propensity score. *J. Am. Stat. Assoc.* 99(467):854–66
Imbens GW. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87:706–10
Imbens GW. 2004. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Stat.* 86(1):4–29
Imbens GW, Wooldridge JM. 2008. *Recent developments in the econometrics of program evaluation*. NBER Work. Pap. No. 14251
Kempthorne O. 1952. *The Design and Analysis of Experiments*. New York: Wiley
Kempthorne O. 1955. The randomization theory of experimental inference. *J. Am. Stat. Assoc.* 50:495–97
Korkeamäki O, Uuistalo R. 2009. Employment and wage effects of a payroll-tax cut—evidence from a regional experiment. *Int. Tax Public Finance.* In press
Kullback S, Leibler RA. 1951. On information and sufficiency. *Ann. Math. Stat.* 22:79–86
Lee DS. 2008. Randomized experiments from non-random selection in U.S. House elections. *J. Econ.* 142(2):675–97
Lehmann EL. 1993. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *J. Am. Stat. Assoc.* 88:1242–49
Lenz GS, Ladd JM. 2006. Exploiting a rare shift in communication flows: media effects in the 1997 British election. **http://sekhon.berkeley.edu/causalinf/papers/LaddLenzBritish.pdf**
Lewis DK. 1973. *Counterfactuals*. Cambridge, MA: Harvard Univ. Press
Mauldon J, Malvin J, Stiles J, Nicosia N, Seto E. 2000. *Impact of California's Cal-Learn demonstration project: final report*. UC DATA Archive and Techn. Assist.
McCarthy MD. 1939. On the application of the *z*-test to randomized blocks. *Ann. Math. Stat.* 10:495–97
Mebane WRJ, Sekhon JS. 2009. Genetic optimization using derivatives: the RGENOUD package for R. *J. Stat. Softw.* In press
Middleton JA. 2008. Bias of the regression estimator for experiments using clustered random assignment. *Stat. Probability Lett.* 78(16):2654–59
Mitchell AFS, Krzanowski WJ. 1985. The Mahalanobis distance and elliptic distributions. *Biometrika* 72(2):464–67
Mitchell AFS, Krzanowski WJ. 1989. Amendments and corrections: the Mahalanobis distance and elliptic distributions. *Biometrika* 76(2):407
Morgan SL, Harding DJ. 2006. Matching estimators of causal effects: prospects and pitfalls in theory and practice. *Sociol. Methods Res.* 35(1):3–60
Morgan SL, Winship C. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge Univ. Press
Mueller FH. 1945. Review of: "experimental sociology: a study in method" by Ernest Greenwood. *Am. Catholic Sociol. Rev.* 6(3):185–86

Neyman J. 1990 (1923). On the application of probability theory to agricultural experiments essay on principles. Sec. 9 *Stat. Sci.* 5(4):465–72. Transl. DM Dabrowska, TP Speed
Pearl J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge Univ. Press
Pitman EJG. 1937. Significance tests which can be applied to samples from any populations. iii. The analysis of variance test. *Biometrika* 29:322–35
Preece DA. 1990. R.A. Fisher and experimental design: a review. *Biometrics* 46(4):925–35
Raessler S, Rubin DB. 2005. Complications when using nonrandomized job training data to draw causal inferences. *Proc. Int. Stat. Inst.*
Reid C. 1982. *Neyman from Life*. New York: Springer
Rosenbaum PR. 1991. A characterization of optimal designs for observational studies. *J. R. Stat. Soc. Ser. B* 53(3):597–610
Rosenbaum PR. 2002. *Observational Studies*. New York: Springer-Verlag. 2nd ed.
Rosenbaum PR. 2005. Heterogeneity and causality: unit heterogeneity and design sensitivity in observational studies. *Am. Stat.* 59:147–52
Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55
Rosenbaum PR, Rubin DB. 1984. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* 79(387):516–24
Rosenbaum PR, Rubin DB. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am. Stat.* 39(1):33–38
Rosenzweig MR, Wolpin KI. 2000. Natural "natural experiments" in economics. *J. Econ. Lit.* 38:827–74
Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66:688–701
Rubin DB. 1976a. Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics* 32(1):109–20
Rubin DB. 1976b. Multivariate matching methods that are equal percent bias reducing, II: Maximums on bias reduction for fixed sample sizes. *Biometrics* 32(1):121–32
Rubin DB. 1977. Assignment to a treatment group on the basis of a covariate. *J. Educ. Stat.* 2:1–26
Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6(1):34–58
Rubin DB. 1979. Using multivariate sampling and regression adjustment to control bias in observational studies. *J. Am. Stat. Assoc.* 74:318–28
Rubin DB. 1980. Bias reduction using Mahalanobis-metric matching. *Biometrics* 36(2):293–98
Rubin DB. 1990. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Stat. Sci.* 5(4):472–80
Rubin DB. 1997. Estimating causal effects from large data sets using propensity scores. *Ann. Int. Med.* 127(8S):757–63
Rubin DB. 2001. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* 2(1):169–88
Rubin DB. 2006. *Matched Sampling for Causal Effects*. New York: Cambridge Univ. Press
Rubin DB. 2008. For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* 2(3):808–40
Rubin DB, Stuart EA. 2006. Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *Ann. Stat.* 34(4):1814–26
Rubin DB, Thomas N. 1992. Affinely invariant matching methods with ellipsoidal distributions. *Ann. Stat.* 20(2):1079–93
Scheffé H. 1956. Alternative models for the analysis of variance. *Ann. Math. Stat.* 27:251–71
Sekhon JS. 2004. *The varying role of voter information across democratic societies*. Work. Pap., Dep. Polit. Sci., Univ. Calif. Berkeley
Sekhon JS. 2006. *Alternative balance metrics for bias reduction in matching methods for causal inference*. Work. Pap., Dep. Polit. Sci., Univ. Calif. Berkeley
Sekhon JS. 2009. Matching: multivariate and propensity score matching with automated balance search. *J. Stat. Softw.* In press. Computer program available at **http://sekhon.berkeley.edu/matching/**
Sekhon JS, Grieve R. 2008. *A new non-parametric matching method for bias adjustment with applications to economic evaluations*. iHEA 2007 6th World Congr., Explorations in Health Econ. Pap.

Sekhon JS, Mebane WR Jr. 1998. Genetic optimization using derivatives: theory and application nonlinear models. *Polit. Anal.* 7:189–203

Simon H. 1953. Causal ordering and identifiability. In *Studies in Econometric Method*, ed. WC Hood, T Koopmans, pp. 49–74. New York: Wiley

Smith HL. 1997. Matching with multiple controls to estimate treatment effects in observational studies. *Sociol. Methodol.* 27:305–53

Smith J, Todd P. 2005a. Does matching overcome LaLonde's critique of nonexperimental estimators? *J. Econ.* 125(1–2):305–53

Smith J, Todd P. 2005b. Rejoinder. *J. Econ.* 125(1–2):365–75

Smith JA, Todd PE. 2001. Reconciling conflicting evidence on the performance of propensity score matching methods. *AEA Pap. Proc.* 91(2):112–18

Snow J. 1855. *On the Mode of Communication of Cholera*. London: John Churchill. 2nd ed.

Sommer A, Zeger SL. 1991. On estimating efficacy from clinical trials. *Stat. Med.* 10(1):45–52

Speed TP. 1990. Introductory remarks on Neyman (1923). *Stat. Sci.* 5(4):463–64

Stigler SM. 1990. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Belknap

Stock JH, Trebbi F. 2003. Who invented instrumental variable regression? *J. Econ. Perspect.* 17(3):177–94

Thistlethwaite DL, Campbell DT. 1960. Regression-discontinuity analysis: an alternative to the ex post facto experiment. *J. Educ. Psychol.* 51(6):309–17

Vinten-Johansen P, Brody H, Paneth N, Rachman S, Rip MR. 2003. *Cholera, Chloroform, and the Science of Medicine: A Life of John Snow*. New York: Oxford Univ. Press

Welch BL. 1937. On the *z*-test in randomized blocks and Latin squares. *Biometrika* 29:21–52

Winship C, Morgan S. 1999. The estimation of causal effects from observational data. *Annu. Rev. Sociol.* 25:659–707

Woo MJ, Reiter JP, Karr AF. 2008. Estimation of propensity scores using generalized additive models. *Stat. Med.* 27:3805–16.

Yule UG. 1899. An investigation into the causes of changes in pauperism in England, chiefly during the last two intercensal decades (Part I). *J. R. Stat. Soc.* 62(2):249–95

Annual Review of
Political Science
Volume 12, 2009

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Political Science* articles may be found at http://polisci.annualreviews.org/