The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods*

Forthcoming in *The Oxford Handbook of Political Methodology*,
Janet Box-Steffensmeier, Henry Brady, and David Collier, eds.

Jasjeet S. Sekhon †

11/16/2007 (15:57)

*I thank Henry Brady, Wendy Tam Cho, David Collier and Rocío Titiunik for valuable comments on earlier drafts, and David Freedman, Walter R. Mebane, Jr., Donald Rubin and Jonathan N. Wand for many valuable discussions on these topics. All errors are my responsibility.

†Associate Professor, Travers Department of Political Science, Survey Research Center, UC Berkeley. sekhon@berkeley.edu, HTTP://sekhon.berkeley.edu/, 510.642.9974.

"Correlation does not imply causation" is one of the most repeated mantras in the social sciences, but its full implications are sobering and often ignored. The Neyman-Rubin model of causal inference helps to clarify some of the issues which arise. In this chapter, the model is briefly described, and some consequences of the model are outlined for both quantitative and qualitative research. The model has radical implications for work in the social sciences given current practices. Matching methods, which are usually motivated by the Neyman-Rubin model, are reviewed and their properties discussed. For example, applied researchers are often surprised to learn that even if the selection on observables assumption is satisfied, the commonly used matching methods will generally make even linear bias worse unless specific and often implausible assumptions are satisfied.

Some of the intuition of matching methods, such as propensity score matching, should be familiar to social scientists because they share many features with Mill's methods, or canons, of inference. Both attempt to find comparable units to compare—i.e., they attempt to increase unit homogeneity. But Mill never intended for his canons to be used in the social sciences because he did not believe that unit homogeneity could be achieved in this field (Sekhon 2004a). Because of its reliance on random assignment and other statistical apparatus, modern concepts of experimental design sharply diverge from Mill's deterministic methods. Modern matching methods adopt Mill's key insights of the importance of unit homogeneity to cases where analysts do not control their units precisely. Matching methods, and related methods such as regression discontinuity, drop observations to make inferences more precise as well as less biased because unit homogeneity can be improved by removing some observations from consideration.[1] Dropping observations is almost anathema to most quantitative researchers, but this intuition is wrong with non-experimental data (Rosenbaum 2005). Case study research methods in the tradition of Mill contrast sharply with statistical methods, and the hunt for necessary and sufficient causes is generally misplaced in the social sciences given the lack of unit homogeneity.

[1]Regression discontinuity is discussed in detail in Chapter 15 (Green and Gerber 2008).

The key probabilistic idea upon which statistical causal inference relies is conditional probability. But when we are making causal inferences, conditional probabilities are not themselves of direct interest. We use conditional probabilities to learn about counterfactuals of interest—e.g., would Jane have voted if someone from the campaign had not gone to her home to encourage her to do so? One has to be careful to establish the relationship between the counterfactuals of interest and the conditional probabilities one has managed to estimate. Researchers too often forget that this relationship must be established by design and instead rely upon statistical models whose assumptions are almost never defended. A regression coefficient is not a causal estimate unless a large set of assumptions are met, and this is no less true of conditional probabilities estimated in other ways such as by matching methods. 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. 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.

1 Neyman-Rubin Causal Model

The Neyman-Rubin framework has become increasingly popular in many fields including statistics (Holland 1986; Rubin 2006, 1974; Rosenbaum 2002), medicine (Christakis and Iwashyna 2003; Rubin 1997), economics (Abadie and Imbens 2006; Galiani, Gertler, and Schargrodsky 2005; Dehejia and Wahba 2002, 1999), political science (Bowers and Hansen 2005; Imai 2005; Sekhon 2004b), sociology (Morgan and Harding 2006; Diprete and Engelhardt 2004; Winship and Morgan 1999; Smith 1997) and even law (Rubin 2001). The framework originated with Neyman's (1923 [1990]) non-parametric model where each unit has two potential outcomes, 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 (2006, 1974), among others, including most notably Cochran (1965; 1953), developed the model into a general framework for causal inference with implications for observational research. Holland (1986) wrote an influential review article which 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 of causal inference (e.g., Brady 2008).

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 set of 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 $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$.[2] Note that in contrast to the usual regression assumptions, the potential outcomes, Y_{i1} and Y_{i1}, are fixed quantities and not random variables.

1.1 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. As the sample size grows, observed and unobserved confounders are balanced across treatment and control groups with arbitrarily

[2]Extensions to the case of multiple discrete treatment are straightforward (e.g., Imbens 2000, Rosenbaum 2002 300–302).

high probability. That is, with random assignment, the distributions of both observed and unobserved variables in both groups are equal in expectation. 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 (Dawid 1979). Hence, for $j = 0, 1$

$$E(Y_{ij} \mid T_i = 1) = E(Y_{ij} \mid T_i = 0) = E(Y_i \mid T_i = j)$$

Therefore, the average treatment effect (ATE) can be estimated by:

$$\begin{aligned} \tau &= E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 0) && (1) \\ &= E(Y_i \mid T_i = 1) - E(Y_i \mid T_i = 0) && (2) \end{aligned}$$

Equation 2 can be estimated consistently in an experimental setting because randomization can ensure that observations in treatment and control groups are exchangeable. Randomization ensures that assignment to treatment will not, in expectation, be associated with the potential outcomes.

Even in an experimental setup, much can go wrong which requires statistical adjustment (e.g., Barnard, Frangakis, Hill, and Rubin 2003). One of the most common problems which arises is the issue of compliance. People who are assigned to treatment may refuse it, and those assigned to control may find some way to receive treatment. When there are compliance issues, Equation 2 then defines the intention-to-treat (ITT) estimand. Although the concept of ITT dates earlier, the phrase probably first appeared in print in Hill (1961, 259). Moving beyond the ITT to estimate the average treatment effect on the treated can be difficult. If the compliance problem is simply that some people assigned to treatment refused it, statistical correction is straightforward and relatively model free. When the compliance problem has a more complicated structure, it is difficult to make progress without making structural assumptions. Statistical corrections for compliance are discussed in detail in Chapter 15 of this volume (Green and Gerber 2008).

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, §2.4). Rubin (1978) calls this "no interference between units" 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 no different versions of treatment. SUTVA is a complicated assumption which is all too often ignored. It is discussed in detail in Chapter 10 (Brady 2008).

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 treatment (Mauldon, Malvin, Stiles, Nicosia, and Seto 2000). Therefore, the experiment probably underestimated the effect of the treatment.

Some researchers erroneously think that the SUTVA assumption is another word for the usual independence assumption made in regression models. A hint of the problem can be seen by noting that 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, an experiment will not yield unbiased estimates of the causal effect of interest. In the usual regression setup, the correct specification assumption (and not the independence assumption) implicitly deals with SUTVA violations. It is assumed that if there are SUTVA violations, we have the correct model for them.

Note that even with randomization, the assumptions of the OLS regression model are not satisfied. Indeed, without further assumptions, the multiple regression estimator is biased, although the bias goes to zero as the sample size increases. And the regression standard errors can be seriously biased, even asymptotically. For details see Freedman (2007a,b). 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. Researchers should be extremely cautious about using multiple regression to adjust experimental data. Unfortunately, there is a tendency to do just that. One supposes that this is yet another sign, as if one more were needed, of how ingrained the regression model is in our quantitative practice.

The only thing stochastic in the Neyman setup 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.

1.2 Observational Data

In an observational setting, unless something special is done, treatment and non-treatment 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):

$$\tau \mid (T = 1) \quad = \quad E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 1). \tag{3}$$

Equation 3 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 X. Following Rosenbaum and Rubin (1983), one can assume that conditional on X, treatment assignment is unconfounded ($\{Y_0, Y_1 \perp\!\!\!\perp T\} \mid X$) and that there is overlap: $0 < Pr(T = 1 \mid X) < 1$.

Together, unconfoundedness and overlap constitute a property known as strong ignorability of assignment which is necessary for identifying the treatment effect. Heckman, Ichimura, Smith, and Todd (1998) shows that for ATT, the unconfoundedness assumption can be weakened to mean independence: $E\left(Y_{ij} \mid T_i, X_i\right) = E\left(Y_{ij} \mid X_i\right)$.[3]

Then, following Rubin (1974, 1977) we obtain

$$E(Y_{ij} \mid X_i, T_i = 1) = E(Y_{ij} \mid X_i, T_i = 0) = E(Y_i \mid X_i, T_i = j). \tag{4}$$

By conditioning on observed covariates, X_i, treatment and control groups are balanced. The average treatment effect for the treated is estimated as

$$\tau \mid (T = 1) \;\; = \;\; E\left\{E(Y_i \mid X_i, T_i = 1) - E(Y_i \mid X_i, T_i = 0) \mid T_i = 1\right\}, \tag{5}$$

where the outer expectation is taken over the distribution of $X_i \mid (T_i = 1)$ which is the distribution of baseline variables in the treated group.

Note that the ATT estimator is changing how individual observations are weighted, and that observations which are 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 which do not overlap with control observations (Crump, Hotz, Imbens, and Mitnik 2006). In such cases, it is unclear exactly the estimand one is estimating because it is no longer ATT as some treatment observations have been dropped along with some control observations.

It is often jarring for people to observe that observations are being dropped because of a lack of covariate overlap. But dropping observations which are outside of common

[3]Also see Abadie and Imbens (2006).

support not only reduces bias but can also reduce the variance of our estimates. This may be counter intuitive, 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 the mean square error. Dropping observations outside of common support and conditioning as in Equation 5 helps to improve unit homogeneity and may actually reduce our variance estimates (Rosenbaum 2005). Moreover, as Rosenbaum (2005) shows, 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.

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 post- treatment? If not, one is unable to even form the relevant questions. The most useful of which may be the one suggested by Dorn (1953, 680) who proposed that the designer of every observational study should ask "[h]ow would the study be conducted if it were possible to do it by controlled experimentation?" This clear question also appears in Cochran's famous Royal Statistical Society discussion paper on the planning of observational studies of human populations (1965). And Dorn's question has become one which researchers in the tradition of the Neyman-Rubin model ask 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. It is fair to say that without answering Dorn's question, one is unsure what the researcher wants to estimate. Since 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 one is unclear as to what precisely 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 which 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 technical discussion which follows on the best way to condition on covariates.

2 Matching Methods

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). The most straightforward and nonparametric way to condition on X is to exactly match on the covariates. This is an old approach going back to at least Fechner (1966 [1860]), the father of psychophysics. This approach fails in finite samples if the dimensionality of X is large and is simply impossible if X contains continuous covariates. Thus, in general, alternative methods must be used.

Two common approaches are propensity score matching (Rosenbaum and Rubin 1983) and multivariate matching based on Mahalanobis distance (Cochran and Rubin 1973; Rubin 1979, 1980). 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[4]) have the property of "equal percent bias reduction" (EPBR) (Rubin 1976a,b; Rubin and Thomas 1992).[5] This property, which is formally defined in Appendix A,

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

[5]The EPBR results of Rubin and Thomas (1992) have been extended by Rubin and Stuart (2005) to the case of discriminant mixtures of proportional ellipsoidally symmetric (DMPES) distributions. This extension

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).

2.1 Mahalanobis and Propensity Score Matching

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

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

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, $md()$.[6] If X consists of more than one continuous variable, multivariate matching estimates contain a bias term which does not asymptotically go to zero at $\sqrt{n}$ (Abadie and Imbens 2006).

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 and Rubin 1983).

Let $e(X_i) \equiv Pr(T_i = 1 \mid X_i) = E(T_i \mid X_i)$, defining $e(X_i)$ to be the propensity score. Given $0 < Pr(T_i \mid X_i) < 1$ and that $Pr(T_1, T_2, \cdots T_N \mid X_1, X_2, \cdots X_N) = \Pi_{i=1}^{N} e(X_i)^{T_i} (1 -$

is important, but it is restricted to a limited set of mixtures. See Appendix A.

[6] One can do matching with replacement or without. Alternatively one can do optimal full matching (Hansen 2004; Rosenbaum 1991) instead of the greedy matching. But this decision is a separate one from the choice of a distance metric.

[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 and Thompson 1952).

$e(X_i))^{(1-T_i)}$, then as Rosenbaum and Rubin (1983) prove,

$$\tau \mid (T = 1) \;\; = \;\; E\left\{E(Y_i \mid e(X_i), T_i = 1) - E(Y_i \mid e(X_i), T_i = 0) \mid T_i = 1\right\},$$

where the outer expectation is taken over the distribution of $e(X_i) \mid (T_i = 1)$. Since the propensity score is generally unknown, it must be estimated.

Propensity score matching involves matching each treated unit to the nearest control unit on the unidimensional metric of the propensity score vector. 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. Moreover, the linear predictor is often more nearly normally distributed which is of some importance given the EPBR results if the propensity score is matched along with other covariates.

Mahalanobis distance and propensity score matching can be combined in various ways (Rubin 2001; Rosenbaum and Rubin 1985). It is useful to combine the propensity score with Mahalanobis distance matching because propensity score matching is particularly good at minimizing the discrepancy along the propensity score and Mahalanobis distance is particularly good at minimizing the distance between individual coordinates of X (orthogonal to the propensity score) (Rosenbaum and Rubin 1985).

A significant shortcoming of common matching methods, such as Mahalanobis distance and propensity score matching, is that they may (and in practice, frequently do) make balance worse across measured potential confounders. These methods may make balance worse, in practice, even if covariates are distributed ellipsoidally symmetric, because EPBR is a property that obtains 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 covariance matrix

of X, can be improved upon.[8] Moreover, if covariates are neither ellipsoidally symmetric nor are mixtures of DMPES distributions, propensity score matching has good theoretical 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 in a given substantive problem it may not be desirable. This can arise if 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, reducing bias in X_1 will be more important than X_2.

Given these limitations, it may be desirable to use a matching method which algorithmically imposes certain properties when the EPBR property does not hold. Genetic Matching does just that.

2.2 Genetic Matching

Sekhon (2007) and Diamond and Sekhon (2005) propose a matching algorithm, Genetic Matching, which maximizes the balance of observed covariates between treated and control groups. Genetic Matching is a generalization of propensity score and Mahalanobis distance matching, and it has been used by a variety of researchers (e.g., Bonney and Minozzi 2007; Brady and Hui 2006; Gilligan and Sergenti 2006; Gordon and Huber 2007; Herron and Wand forthcoming; Morgan and Harding 2006; Lenz and Ladd 2006; Park 2006; Raessler and Rubin 2005). The algorithm uses a genetic algorithm (Mebane and Sekhon 1998; Sekhon and Mebane 1998) to optimize balance as much as possible given the data. The method 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 and Sekhon (2005) use this algorithm to show that the long running debate between Dehejia and Wahba

[8]For justifications of Mahalanobis distance based on distributional considerations see Mitchell and Krzanowski (1985, 1989).

(2002; 1997; 1999; Dehejia 2005) and Smith and Todd (2005b,a, 2001) is largely a result of researchers using models which do not produce good balance—even if some of the models get close by chance to the experimental benchmark of interest. They show that Genetic Matching is able to quickly find good balance and reliably recover the experimental benchmark.

The idea underlying the Genetic Matching algorithm is that if Mahalanobis distance is not optimal for achieving balance in a given dataset, one should be able to search over the space of distance metrics and find something better. One way of generalizing the Mahalanobis metric is to include an additional weight matrix:

$$d(X_i, X_j) \quad = \left\{ (X_i - X_j)' \left(S^{-1/2}\right)' W S^{-1/2} (X_i - X_j) \right\}^{\frac{1}{2}}$$

where W is a $k \times k$ positive definite weight matrix and $S^{1/2}$ is the Cholesky decomposition of S which is the variance-covariance matrix of X.[9]

Note that if one has a good propensity score model, one should include it as one of the covariates in Genetic Matching. If this is done, both propensity score matching and Mahalanobis matching can be considered special limiting cases of Genetic Matching. If the propensity score contains all of the relevant information in a given sample, the other variables will be given zero weight.[10] And Genetic Matching will converge to Mahalanobis distance if that proves to be the appropriate distance measure.

Genetic Matching is an affinely invariant matching algorithm that uses the distance measure $d()$, in which all elements of W are zero except down the main diagonal. The main diagonal consists of k parameters which must be chosen. Note that if each of these k parameters are set equal to 1, $d()$ is the same as Mahalanobis distance.

The choice of setting the non-diagonal elements of W to zero is made for reasons of computational power alone. The optimization problem grows exponentially with the number

[9]The Cholesky decomposition is parameterized such that $S = LL'$, $S^{1/2} = L$. In other words, L is a lower triangular matrix with positive diagonal elements.

[10]Technically, the other variables will be given weights just large enough to ensure that the weight matrix is positive definite.

of free parameters. It is important that the problem be parameterized so as to limit the number of parameters which must be estimated.

This leaves the problem of how to choose the free elements of W. Many loss criteria recommend themselves, and many can be used with the software which implements Genetic Matching.[11] By default, cumulative probability distribution functions of a variety of standardized statistics are used as balance metrics and are optimized without limit. The default standardized statistics are paired t-tests and bootstrapped nonparametric Kolmogorov-Smirnov tests (Abadie 2002).

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. Alternatively, one may choose to minimize some descriptive measure of imbalance such as the maximum gap in the standardized empirical-QQ plots across the covariates. This would correspond to minimizing the D statistic of the Kolmogorov-Smirnov test.

Conceptually, the algorithm attempts to minimize the largest observed covariate discrepancy at every step. This is accomplished by maximizing the smallest p-value at each step.[12] Because Genetic Matching is minimizing the maximum discrepancy observed at each step, it is minimizing the infinity norm. This property holds even when, because of the distribution of X, the EPBR property does not hold. Therefore, if an analyst is concerned that matching may increase the bias in some linear combination of X even if the means are reduced, Genetic Matching allows the analyst to put in the loss function all of the linear combinations of X which may be of concern. Indeed, any nonlinear function of X can also be included in the loss function, which would ensure that bias in some nonlinear functions of X is not made inordinately large by matching.

[11] See `http://sekhon.berkeley.edu/matching`.

[12] More precisely, lexical optimization will be done: all of the balance statistics will be sorted from the most discrepant to the least and weights will be picked which minimize the maximum discrepancy. If multiple sets of weights result in the same maximum discrepancy, then the second largest discrepancy is examined to choose the best weights. The processes continues iteratively until ties are broken.

The default Genetic Matching loss function does allow for imbalance in functions of X to worsen as long as the maximum discrepancy is reduced. This default behavior can be altered by the analyst. It is important that the maximum discrepancy be small—i.e., that the smallest p-value be large. The p-values conventionally understood to signal balance (e.g., 0.10), may be too low to produce reliable estimates. After Genetic Matching optimization, the p-values from these balance tests cannot be interpreted as true probabilities because of standard pre-test problems, but they remain useful measures of balance. Also, we are interested in maximizing the balance in the current sample so a hypothesis test for balance is inappropriate.

The optimization problem described above is difficult and irregular, and the genetic algorithm implemented in the *R* `rgenoud` package (Mebane and Sekhon 1998) is used to conduct the optimization. Details of the algorithm are provided in Sekhon and Mebane (1998).

Genetic Matching is shown to have better properties than the usual alternative matching methods both when the EPBR property holds and when it does not (Sekhon 2007; Diamond and Sekhon 2005). Even when the EPBR property holds and the mapping from X to Y is linear, Genetic Matching has better efficiency—i.e., lower mean square error (MSE)—in finite samples. When the EPBR property does not hold as it generally does not, Genetic Matching retains appealing properties and the differences in performance between Genetic Matching and the other matching methods can become substantial both in terms of bias and MSE reduction. In short, at the expense of computer time, Genetic Matching dominates the other matching methods in terms of MSE when assumptions required for EPBR hold and, even more so, when they do not.

Genetic Matching is able to retain good properties even when EPBR does not hold because a set of constraints is imposed by the loss function optimized by the genetic algorithm. The loss function depends on a large number of functions of covariate imbalance across matched treatment and control groups. Given these measures, Genetic Matching will

optimize covariate balance.

3 Case Study Research Methods

Matching designs have long been used by social scientists conducting qualitative research methods. But case study matching methods often rely on the assumption that the relationships between the variables of interest are deterministic. This is unfortunate because failure to heed the lessons of statistical inference often leads to serious inferential errors, some of which are easy to avoid. The canonical example of deterministic matching designs methods is the set of rules (canons) of inductive inference formalized by John Stuart Mill in his *A System of Logic* (1872).

The "most similar" and the "most different" research designs, which are often used in comparative politics, are variants of Mill's methods (Przeworski and Teune 1970). As such, Mill's methods have been used by generations of social science researchers (Cohen and Nagel 1934), but they contrast sharply with statistical methods. These methods do not lead to valid inductive inferences unless a number of very special assumptions hold. Some researchers seem to be either unaware or unconvinced of these methodological difficulties even though the acknowledged originator of the methods, Mill himself, clearly described many of their limitations.

These canonical qualitative methods of causal inference are only valid when the hypothesized relationship between the cause and effect of interest is *unique* and *deterministic*. These two conditions imply other conditions such as the absence of measurement error which would cease to make the hypothesized causal relationship deterministic as least as we observe it. These assumptions are strict, and they strongly restrict the applicability of the methods. When these methods of inductive inference are not applicable, conditional probabilities should be used to compare the relevant counterfactuals.[13]

[13]Needless-to-say, although Mill was familiar with the work of Laplace and other 19th century statisticians, by today's standards his understanding of estimation and hypothesis testing was simplistic and often

For these methods to lead to valid inferences there must be only one possible cause of the effect of interest, the relationship between cause and effect must be deterministic, and there must be no measurement error. If these assumptions are to be relaxed, random factors must be accounted for. Because of these random factors, statistical and probabilistic methods of inference are necessary.

To appreciate how serious these limitations are, consider the case of benchmarking statistical software on modern computer systems—for details see Sekhon (2006). Such computers are Turing machines hence they are deterministic systems where everything a computer does is in theory observable. To put it another way, your random number generator is not really random. Your pseudorandom numbers are the result of a deterministic algorithm. But notwithstanding the deterministic nature of a computer, methods like those proposed by qualitative researchers for making inferences with deterministic systems are not used in the benchmarking literature. When benchmarking, it is common to match on (and hence eliminate) as many confounders as possible and to report measures of uncertainty and statistical hypothesis tests. Since computers are deterministic, the remaining uncertainty must come from confounders—as opposed to sampling error—which could in theory be observed and hence eliminated. But the system is considered to be so complex that most benchmarking exercises resort to statistical measures of association. Thus, even in this setting where we know we are dealing with a deterministic system, benchmarking exercises rely on statistical measures because of the complexity involved. Certainly society is more complex than a computer and our social measurements are more prone to error than those of computers.

3.1 Mill's Methods and Conditional Probabilities

Since the application of the five methods Mill discusses has a long history in the social sciences, I am hardly the first to criticize the use of these methods in all but very special

erroneous. He did, however, understand that if one wants to make valid empirical inferences, one needs to obtain and compare conditional probabilities when there may be more than one possible cause of an effect or when the causal relationship is complicated by interaction effects.

circumstances. For example, Robinson, who is well known in political science for his work on the ecological inference problem,[14] also criticized the use of Mill-type methods of analytic induction in the social sciences (Robinson 1951). Robinson's critique did not, however, focus on conditional probabilities nor did he observe that Mill himself railed against the exact use to which his methods have been put. Many other critics will be encountered in the course of our discussion.

Przeworski and Teune, in an influential book, advocate the use of what they call the "most similar" design and the "most different" design (Przeworski and Teune 1970). These designs are variations on Mill's methods. The first is a version of Mill's Method of Agreement, and the second is a *weak* version of Mill's Method of Difference. Although the Przeworski and Teune volume is over 30 years old, their argument continues to be influential. For example, Ragin, Berg-Schlosser, and de Meur in a recent review of qualitative methods make direct supportive references to both Mill's methods and Przeworski and Teune's formulations (Ragin et al. 1996). However, even when authors such as Ragin et al. recognize that Mill's methods need to be altered for use in the social sciences, usually follow neither the advice of quantitative methodologists nor Mill's own advice regarding the use of conditional probabilities.[15]

Mill described his views on scientific investigations in *A System of Logic Ratiocinative and Inductive*, first published in 1843.[16] In an often cited chapter (bk. III, ch. 8), Mill formulates five guiding methods of induction: the Method of Agreement, the Method of Difference, the Double Method of Agreement and Difference (also known as the Indirect Method of Difference), the Method of Residues, and the Method of Concomitant Variations. These methods are often counted to be only four because the Double Method of Agreement and Difference may be considered to be just a derivative of the first two methods. This is

[14]Ecological inferences are inferences about individual behavior which are based on data of group behavior, called aggregate or ecological data.

[15]For details on the relationship between qualitative comparative analysis and standard regression see Seawright (2004).

[16]For all page referencing I have used a reprint of the eighth edition of *A System of Logic Ratiocinative and Inductive*, first published in 1872. The eighth edition was the last printed in Mill's lifetime. The eighth and third editions were especially revised and supplemented with new material.

a mistake because it obscures the tremendous difference between the combined method or what Mill calls the Indirect Method of Difference and the Direct Method of Difference (Mill 1872, 259). Both the Method of Agreement and the Indirect Method of Difference, which is actually the Method of Agreement applied twice, are limited and require the machinery of probability in order to take chance into account when considering cases where the number of causes may be greater than one or where there may be interactions between the causes (Mill 1872, 344). Other factors not well explored by Mill, such as measurement error, lead to the same conclusion (Lieberson 1991). The Direct Method of Difference is almost entirely limited to the experimental setting. And even in the case of the Direct Method of Difference, chance must be taken into account in the presence of measurement error or if there are interactions between causes which lead to probabilistic relationships between a cause, A, and its effect, a.

Next, we review Mill's first three canons and show the importance of taking chance into account and comparing conditional probabilities when chance variations cannot be ignored.

3.1.1 First Canon: Method of Agreement

> "If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree is the cause (or effect) of the given phenomenon" (Mill 1872, 255).

Assume that the *possible* causes, i.e., antecedents, under consideration are denoted by A, B, C, D, E, and the effect we are interested in is denoted by a.[17] An antecedent may be comprised of more than one constituent event or condition. For example, permanganate ion with oxalic acid forms carbon dioxide (and manganous ion). Separately, neither permanganate ion nor oxalic acid will produce carbon dioxide, but if combined, they will. In this example, A may be defined as the presence of both permanganate ion and oxalic acid.

[17]Following Mill's usage, my usage of the word "antecedent" is synonymous with "possible cause." Neither Mill nor I intend to imply that events *must* be ordered in time to be causally related.

Let us further assume that we observe two instances and in the first we observe the antecedents A, B, C, and in the second we observe the antecedents A, D, E. If we also observe the effect, a, in both cases, we would say, following Mill's Method of Agreement, that A is the cause of a. We conclude this because A was the only antecedent which occurred in both observations—i.e., the observations agree on the presence of antecedent A. This method has eliminated antecedents B, C, D, E as possible causes of a. Using this method, we endeavor to obtain observations which agree in the effect, a, and the supposed cause, A, but differ in the presence of other antecedents.

3.1.2 Second Canon: Method of Difference

> "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon" (Mill 1872, 256).

In the Method of Difference we require, contrary to the Method of Agreement, observations resembling one another in every other respect, but differing in the presence or absence of the antecedent we conjecture to be the true cause of a. If our object is to discover the effects of an antecedent A, we must introduce A into some set of circumstances we consider relevant, such as B, C, and having noted the effects produced, compare them with the effects of the remaining circumstances B, C, when A is absent. If the effect of A, B, C is a, b, c, and the effect of B, C is b, c, it is evident, under this argument, that the cause of a is A.

Both of these methods are based on a process of elimination. This process has been understood since Francis Bacon to be a centerpiece of inductive reasoning (Pledge 1939). The Method of Agreement is supported by the argument that whatever can be eliminated is not connected with the phenomenon of interest, a. The Method of Difference is supported by the argument that whatever cannot be eliminated is connected with the phenomenon by

a law. Because both methods are based on the process of elimination, they are deterministic in nature. For if even one case is observed where effect a occurs without the presence of antecedent A, we would eliminate antecedent A from causal consideration.

Mill asserts that the Method of Difference is commonly used in experimental science while the Method of Agreement, which is substantially weaker, is employed when experimentation is impossible (Mill 1872, 256). The Method of Difference is Mill's attempt to describe the inductive logic of experimental design. And the method takes into account two of the key features of experimental design, the first being the presence of a manipulation (treatment) and the second a comparison between two states of the world which are in Mill's case exactly alike aside from the presence of the antecedent of interest.[18] The method also incorporates the notion of a relative causal effect. The effect of antecedent A is measured relative to the effect observed in the most similar world without A. The two states of the world we are considering only differ in the presence or absence of A.

The Method of Difference only accurately describes a small subset of experiments. The method is too restrictive even if the relationship between the antecedent A and effect a were to be deterministic. Today we would say that the control group B, C and the group with the intervention A, B, C need not be *exactly* alike (aside from the presence or absence of A). It would be fantastic if the two groups were exactly alike, but such a situation is not only extremely difficult to find but also not necessary. Some laboratory experiments are based on this strong assumption, but a more common assumption, and one which brings in statistical concerns, is that observations in both groups are *balanced* before our intervention. That is, before we apply the treatment, the distributions of both observed and unobserved variables in both groups are equal. For example, if group A is the southern states in the United States and group B is the northern states, the two groups are not balanced. The distribution of a

[18]The requirement of a manipulation by the researcher has troubled many philosophers of science. But the claim is not that causality requires a human manipulation, but only that if we wish to measure the effect of a given antecedent we gain much if we are able to manipulate the antecedent. For example, manipulation of the antecedent of interest allows us to be confident that the antecedent caused the effect and not the other way around—see Brady (2008).

long list of variables is different between the two groups.

Random assignment of treatment ensures, if the sample size is large and if other assumptions are met, that the control and treatment groups are balanced even on unobserved variables.[19] Random assignment ensures that the treatment is uncorrelated with all baseline variables[20] whether we can observe them or not.[21]

Because of its reliance on random assignment, modern concepts of experimental design sharply diverge from Mill's deterministic model. The two groups are not exactly alike in baseline characteristics (as they would have to be in a deterministic setup), but, instead, their baseline characteristics have the same distribution. And consequently the baseline variables are uncorrelated with whether a particular unit received treatment or not.

When the balance assumption is satisfied, a modern experimenter estimates the relative causal effect by comparing the conditional probability of some outcome given the treatment minus the conditional probability of the outcome given that the treatment was not received. In the canonical experimental setting, conditional probabilities can be directly interpreted as causal effects.

In the penultimate section of this chapter, I discuss the complications which arise in using conditional probabilities to make causal inferences when randomization of treatment is not possible. With observational data (i.e., data found in nature and not a product of experimental manipulation), many complications arise which prevent conditional probabilities from being directly interpreted as estimates of causal effects. Problems also often arise with experiments which prevent the simple conditional probabilities from being interpreted as relative causal effects. School voucher experiments are a good example.[22] But the prob-

[19]Aside from a large sample size, experiments need to also meet a number of other conditions. See Campbell and Stanley (1966) for an overview particularly relevant for the social sciences. An important problem with experiments dealing with human beings is the issue of compliance. Full compliance implies that every person assigned to treatment actually receives the treatment and every person assigned to control does not. Fortunately, if noncompliance is an issue, there are a number of possible corrections which make few and reasonable assumptions—see Barnard et al. (2003).

[20]Baseline variables are the variables observed before treatment is applied.

[21]More formally, random assignment results in the treatment being stochastically independent of all baseline variables as long as the sample size is large and other assumptions are satisfied.

[22]Barnard et al. (2003) discuss in detail a broken school voucher experiment and a correction using strat-

lems are more serious with observational data where neither a manipulation nor balance are present.[23]

One of the continuing appeals of deterministic methods for case study researchers is the power of the methods. For example, Mill's Method of Difference can determine causality with only two observations. This power can only be obtained by assuming that the observation with the antecedent of interest, A, B, C and the one without, B, C are exactly alike accept for the manipulation of A, and by assuming deterministic causation and the absence of measurement error and interactions among antecedents. This power makes deterministic methods alluring for case study researchers, who generally don't have many observations. Once probabilistic factors are introduced, larger numbers of observations are required to make useful inferences. Because of the power of deterministic methods, social scientists with a small number of observations are tempted to rely on Mill's methods. Because these researchers cannot conduct experiments, they largely rely on the Method of Agreement, which we have discussed, and Mill's third canon.

3.1.3 Third Canon: Indirect Method of Difference

> "If two or more instances in which the phenomenon occurs have only one circumstance in common, while two or more instances in which it does not occur have nothing in common save the absence of that circumstance, the circumstance in which alone the two sets of instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon" (Mill 1872, 259).

This method arises by a "double employment of the Method of Agreement" (Mill 1872, 258). If we observe a set of observations in all of which we observe a and note that they have no antecedent in common but A, by the Method of Agreement we have evidence that A is

ification.

[23] In an experiment much can go wrong (e.g., compliance and missing data problems), but the fact that there was a manipulation can be very helpful in correcting the problems—Barnard et al. (2003). Corrections are more problematic in the absence of an experimental manipulation because additional assumptions are required.

the cause of the effect a. Ideally, we would then perform an experiment where we manipulate A to see if the effect a is present when the antecedent A is absent. When we cannot conduct such an experiment, we can instead use the Method of Agreement again. Suppose, we can find another set of observations in which neither the antecedent A nor the effect a occur. We may now conclude, by use of the Indirect Method of Difference, that A is the cause of a. Thus, by twice using the Method of Agreement we may hope to establish both the positive and negative instance which the Method of Difference requires. However, this double use of the Method of Agreement is clearly inferior. The Indirect Method of Difference cannot fulfill the requirements of the Direct Method of Difference. For, "the requisitions of the Method of Difference are not satisfied unless we can be quite sure either that the instances affirmative of a agree in no antecedents whatever but A, or that the instances negative of a agree in nothing but the negation of A" (Mill 1872, 259). In other words, the Direct Method of Difference is the superior method because it entails a strong manipulation: we manipulate the antecedents so that we can remove the suspected cause, A, and then put it back at will, without disturbing the balance of what may lead to a. And this manipulation ensures that the only difference in the antecedents between the two observations is the presence of A or its lack.

Researchers are often unclear about these distinctions between the indirect and direct methods of difference. They often simply state they are using the Method of Difference when they are actually only using the Indirect Method of Difference. For example, Skocpol states that she is using both the Method of Agreement and the "more powerful" Method of Difference when she is only using at best the weaker Method of Agreement twice (Skocpol 1979, 36–37). It is understandable that Skocpol is not able to use the Direct Method of Difference since it would be impossible to manipulate the factors of interest. But it is important to be clear about exactly which method one is using.

Mill discusses two other canons: the Method of Residues (Fourth Canon) and the Method of Concomitant Variations (Fifth Canon). We do not review these canons because they are

not directly relevant to our discussion.

We have so far outlined the three methods of Mill with which we are concerned. We have also shown that when scholars such as Skocpol assert that they are using the Method of Agreement and the Method of Difference (Skocpol 1979, 37), they are actually using the Indirect Method of Difference, and that this is indeed the weaker sibling of the Direct Method of Difference. This weakness would not be of much concern if the phenomena we studied were simple. However, in the social sciences we encounter serious causal complexities.

Mill's methods of inductive inference are valid only if the mapping between antecedents and effects is *unique* and *deterministic* (Mill 1872, 285–299, 344–350). These conditions allow neither for more than one cause for an effect nor for interactions between causes. In other words, if we are interested in effect a, we must assume *a priori* that only one possible cause exists for a and that when a's cause is present, say cause A, the effect, a, must *always* occur. In fact, these two conditions, of uniqueness and determinism, define the set of antecedents we are considering. This implies, for example, that the elements in the set of causes A, B, C, D, E must be able to occur independently of each other. The condition is not that antecedents must be independent in the probabilistic sense of the word, but that any one of the antecedents can occur without necessitating the presence or lack thereof of any of the other antecedents. Otherwise, the possible effects of antecedents are impossible to distinguish by these rules.[24] Generalizations of Mill's methods also suffer from these limitations (Little 1998, 221-223).

The foregoing has a number of implications the most important of which is that for deterministic methods such as Mill's to work there must be no measurement error. For even if there were a deterministic relationship between antecedent A and effect a, if we were able to measure either A or a only with some stochastic error, the resulting observed relationship

[24]Mill's methods have additional limitations which are outside the scope of this discussion. For example, there is a set of conditions, call it Z, which always exists but is unconnected with the phenomenon of interest. For example, the star Sirius is always present (but not always observable) whenever it rains in Boston. Is the star Sirius and its gravitational force causally related to rain in Boston? Significant issues arise from this question which I do not discuss.

would be probabilistic. It would be probabilistic because it would be possible to observe a case in which we mistakenly think we have observed antecedent A (because of measurement error) while not observing a. In such a situation the process of elimination would lead us to conclude that A is not a cause of a.

To my knowledge no modern social scientist argues that the conditions of uniqueness and lack of measurement error hold in the social sciences. However, the question of whether deterministic causation is plausible has a sizable literature.[25] Most of this discussion centers on whether deterministic relationships are possible—i.e., on the ontological status of deterministic causation.[26] Although such discussions can be fruitful, we need not decide the ontological issues in order to make empirical progress. This is fortunate because the ontological issues are at best difficult to resolve and may be impossible to resolve. Even if one concedes that deterministic social associations exist, it is unclear how we would ever learn about them if there are multiple causes with complex interactions or if our measures are noisy. The case of multiple causes and complex interactions among deterministic associations would, to us, look probabilistic in the absence of a theory (and measurements) which accurately accounted for the complicated causal mechanisms—e.g., Little (1998, ch. 11). There appears to be some agreement among qualitative and quantitative researchers that there is "complexity-induced probabilism" (Bennett 1998). Thus, I think it is more fruitful to focus instead on the practical issue of how we learn about causes—i.e., on the epistemological issues related to causality.[27]

Focusing on epistemological issues also helps to avoid some thorny philosophical questions regarding the ontological status of probabilistic notions of causality. For example, if one can accurately estimate the probability distribution of A causing a, does that mean that we can explain any particular occurrence of a? Wesley Salmon, after surveying three prominent theories of probabilistic causality in the mid-1980s, noted that "the primary moral I drew

[25]See Waldner (2002) for an overview.

[26]Ontology is the branch of philosophy concerned with the study of existence itself.

[27]Epistemology is the branch of philosophy concerned with the theory of knowledge, in particular, the nature and derivation of knowledge, its scope and the reliability of claims to knowledge.

was that causal concepts cannot be fully explicated in terms of statistical relationships; in addition, I concluded, we need to appeal to causal processes and causal interactions" (Salmon 1989, 168). I do not think these metaphysical issues ought to concern practicing scientists.

Faced with multiple causes and interactions what is one to do? There are two dominant responses. The first relies on detailed (usually formal) theories which make precise empirical predictions which distinguish between the theories. Such theories are usually tested by laboratory experiments with such strong manipulations and careful controls that one may reasonably claim to have obtained exact balance and the practical absence of measurement error. Such manipulations and controls allow one to use generalizations of the Method of Difference. A large number of theories in physics offer canonical examples of this approach. Deduction plays a prominent role in this approach.[28]

The second response relies on conditional probabilities and counterfactuals. These responses are not mutually exclusive. Economics, for example, is a field which relies heavily on both formal theories and statistical empirical tests. Indeed, unless the proposed formal theories are nearly complete, there will always be a need to take random factors into account. And even the most ambitious formal modeler will no doubt concede that a complete deductive theory of politics is probably impossible. Given that our theories are weak, our causes complex and data noisy, we cannot avoid conditional probabilities. Even researchers sympathetic to finding necessary or sufficient causes are often led to probability given these problems (e.g., Ragin 2000).

4 From Conditional Probabilities to Counterfactuals

Although conditional probability is at the heart of inductive inference, by itself it isn't enough. Underlying conditional probability is a notion of counterfactual inference. It is

[28]Mill places great importance on deduction in the three step process of "induction, ratiocination, and verification" Mill (1872, 304). But on the whole, although the term *ratiocinative* is in the title of Mill's treatise and even appears before the term *inductive*, Mill devotes little space to the issue of deductive reasoning.

possible to have a causal theory that makes no reference to counterfactuals (Brady 2008; Dawid 2000), but counterfactual theories of causality are by far the norm, especially in statistics. The Method of Difference is motivated by a counterfactual notion: I would like to see what happens both with antecedent A and without A. When I use the Method of Difference, I don't conjecture what would happen if A were absent. I remove A and actually see what happens. Implementation of the method obviously depends on a manipulation. Although manipulation is an important component of experimental research, manipulations as precise as those entailed by the Method of Difference are not possible in the social sciences in particular and with field experiments in general.

We have to depend on other means to obtain information about what would occur both if A is present and if A is not. In many fields, a common alternative to the Method of Difference is a randomized experiment. For example, we could either contact Jane to prompt her to vote as part of a turnout study or we could not contact her. But we cannot observe what would happen if we both contacted Jane and if we did not contact Jane—i.e., we cannot observe Jane's behavior both with and without the treatment. If we contact Jane, in order to determine what effect this treatment had on Jane's behavior (i.e, whether she voted or not), we still have to obtain some estimate of the counterfactual in which we did not contact Jane. We could, for example, seek to compare Jane's behavior with someone exactly like Jane whom we did not contact. The reality, however, is that there is no one exactly like Jane with whom we can compare Jane's turnout decision. Instead, in a randomized experiment we obtain a group of people (the larger the better) and we assign treatment to a randomly chosen subset (to contact) and we assign the remainder to the control group (not to be contacted). We then observe the difference in turnout rates between the two groups and we attribute any differences to our treatment.

In principle the process of random assignment results in the observed and unobserved baselines variables of the two groups being balanced.[29] In the simplest setup, individuals in

[29]This occurs with arbitrarily high probability as the sample size grows.

both groups are supposed to be equally likely to receive the treatment, and hence assignment to treatment will not be associated with anything which may also affect one's propensity to vote. In an observational setting, unless something special is done, the treatment and non-treatment groups are almost never balanced.

The core counterfactual motivation is often forgotten when researchers estimate conditional probabilities to make causal inferences. This situation often arises when quantitative scholars attempt to estimate partial effects.[30] On many occasions the researcher estimates a regression and interprets each of the regression coefficients as estimates of causal effects holding all of the other variables in the model constant. For many in the late 19th and early 20th centuries, this was the goal of the use of regression in the social sciences. The regression model was to give the social scientist the control over data which the physicist obtained via precise formal theories and the biologist obtained via experiments. Unfortunately, if one's covariates are correlated with each other (as they almost always are), interpreting regression coefficients to be estimates of partial causal effects is simply asking too much from the model. With correlated covariates, one variable (such as race) does not move independently of other covariates (such as income, education and neighborhood). With such correlations, it is difficult to posit interesting counterfactuals of which a single regression coefficient is a good estimate.

A good example of these issues is offered by the literatures which developed in the aftermath of the 2000 Presidential election. A number of scholars try to estimate the relationship between the race of a voter and uncounted ballots. Ballots are uncounted either because the ballots contain no votes (undervotes) or overvotes (more than the legal number of votes).[31] If one were able to estimate a regression model, for example, which showed that there was no relationship between the race of a voter and her probability of casting uncounted ballots when and only when one controlled for a long list of covariates, it would be unclear

[30] A partial effect is the effect a given antecedent has on the outcome variable net of all the other antecedents—i.e., when all of the other variables "are held constant."

[31] See Herron and Sekhon (2003; 2005) for a review of the literature and relevant empirical analysis.

what one has found. This uncertainty holds even if ecological and a host of other problems are pushed aside because such a regression model may not allow one to answer the counterfactual question of interest—i.e., "if a black voter became white, would this increase or decrease her chance of casting an uncounted ballot?" What does it mean to change a voter from black to white? Given the data, it is not plausible that changing a voter from black to white would have no implications for the individual's income, education or neighborhood of residence. It is difficult to conceptualize a serious counterfactual for which this regression result is relevant. Before any regression is estimated, we know that if we measure enough variables well, the race variable itself in 2000 will be insignificant. But in a world where being black is highly correlated with socioeconomic variables, it is not clear what we learn about the causality of ballot problems from a showing that the race coefficient itself can be made insignificant.

There are no general solutions or methods which ensure that the statistical quantities we estimate provide useful information about the counterfactuals of interest. The solution, which almost always relies on research design and statistical methods, depends on the precise research question under consideration. But all too often the problem is ignored. All too often the regression coefficient itself is considered to be an estimate of the partial causal effect. Estimates of conditional means and probabilities are an important component of establishing causal effects, but are not enough. One has to establish the relationship between the counterfactuals of interest and the conditional probabilities one has managed to estimate.

A large number of other issues are also important when one is examining the quality of the conditional probabilities one has estimated. A prominent example is the extent to which one can combine a given collection of observations. The combining of observations which are actually rather different is one of the standard objections to statistical analysis. But the question of when and how one can legitimately combine observations is and has long been one of the central research questions in statistics. In fact, the original purpose of least squares was to give astronomers a way of combining and weighting their discrepant

observations in order to obtain better estimates of the locations and motions of celestial objects (Stigler 1986). Generally used methods, such as robust estimation, still require that the model for combining observations is correct for most of the sample under consideration so they do not get to the heart of the problem (e.g., Bartels 1996; Mebane and Sekhon 2004). This is a subject that political scientists need to pay more attention to.

5 Discussion

This chapter has by no means offered a complete discussion of causality and all one has to do in order to demonstrate a causal relationship. There is much more to this than just conditional probabilities and even counterfactuals. For example, it is often important to find the causal mechanism at work, in the sense of understanding the sequence of events which lead from A to a. And I agree with qualitative researchers that case studies are particularly helpful in learning about such mechanisms. Process tracing is often cited as being particularly useful in this regard.[32]

The importance of searching for causal mechanisms is often overestimated by political scientists and this sometimes leads to an underestimate of the importance of comparing conditional probabilities. We do not need to have much or any knowledge about mechanisms in order to know that a causal relationship exists. For example, by the use of rudimentary experiments, aspirin has been known to help with pain since Felix Hoffmann synthesized a stable form of acetylsalicylic acid in 1897. In fact, the bark and leaves of the willow tree (rich in the substance called salicin) have been known to help alleviate pain at least since the time of Hippocrates. But the causal mechanism by which aspirin alleviates pain was a mystery until recently. Only in 1971 did John Vane discover aspirin's biological mechanism of action.[33] And even now, although we know how it crosses the blood-brain barrier, we have little idea

[32]Process tracing is the enterprise of using narrative and other qualitative methods to determine the mechanisms by which a particular antecedent produces its effects—see George and McKeown (1985).

[33]He was awarded the 1982 Nobel Prize for Medicine for this discovery.

how the chemical changes in the brain due to aspirin get translated into the conscious feeling of pain relief—after all, the mind-body problem has not been solved. But knowledge of causal mechanisms is important and useful and no causal account can be considered complete without a mechanism being demonstrated or at the very least hypothesized.

The search for causal mechanisms is probably especially useful when working with observational data. But it is still not necessary. In the case of the causal relationship between smoking and cancer, human experiments were not possible yet most (but not all) neutral researchers were convinced of the causal relationship well before the biological mechanisms were known.[34]

In clinical medicine case studies continue to contribute valuable knowledge even though large-N statistical research dominates. Although the coexistence is sometimes uneasy, as noted by the rise of clinical outcomes research, it is nevertheless extremely fruitful and more cooperative than the relationship in political science.[35] One reason for this is that in clinical medicine, researchers reporting cases more readily acknowledge that the statistical framework helps to provide information about when and where cases are informative (Vandenbroucke 2001). Cases can be highly informative when our understanding of the phenomena of interest is very poor, because then we can learn a great deal from a few observations. On the other hand, when our understanding is generally very good, a few cases which combine a set of circumstances that we believed could not exist or, more realistically, were believed to be highly unlikely can alert us to overlooked phenomena. Some observations are more important than others and there sometimes are "critical cases" (Eckstein 1975). This point is not new to qualitative methodologists because there is an implicit (and all too rarely

[34]R. A. Fisher, one of the fathers of modern statistics and the experimental method, was a notable exception. Without the manipulation offered by an experiment, he remained skeptical. He hypothesized that genetic factors could cause people to both smoke and get cancer, and hence there need not be any causal relationship between smoking and cancer (Fisher 1958a,b).

[35]Returning to the aspirin example, it is interesting to note that Lawrence Craven, a general practitioner, noticed in 1948 that the 400 men he had prescribed aspirin to did not suffer *any* heart attacks. But it was not until 1985 that the U.S. Food and Drug Administration approved the use of aspirin for the purposes of reducing the risk of heart attack. And in 1988 the Physicians' Health Study, a randomized, double-blind, placebo-controlled trial of apparently healthy men, was stopped early because the effectiveness of aspirin had finally been demonstrated (Steering Committee of the Physicians' Health Study Research Group 1989).

explicit) Bayesianism in their discussion of the relative importance of cases (George and McKeown 1985; McKeown 1999). If one only has a few observations, it is more important than otherwise to pay careful attention to the existing state of knowledge when selecting cases and when deciding how informative they are. In general, as our understanding of an issue improves, individual cases become less important.

The logical fallacy of *cum hoc ergo propter hoc* ("with this, therefore because of this") is committed by social scientists as a matter of course. Looking over a random sample of quantitative articles in the *APSR* over the past 30 years, there appears to be no decline in articles which commit this fallacy. The fallacy is now more often committed in a multivariate sense with the use of multiple regression as opposed to correlations or crosstabs. But that does not avoid the problem.

Historically, the matching literature, like much of statistics, has been limited by computational power. What is possible with matching today is nothing like what was possible in 1970 let alone during Mill's time. Not so long ago estimating a logistic regression, the common way today to estimate the propensity score, was prohibitive for all but the smallest of datasets. Today, as we have seen, we can apply machine learning algorithms to the matching problem. These technical innovations will continue, but history teaches us to be cautious about what the technical advances will bring us. Without a greater focus on experimental research and rigorous observational designs, it is unclear what substantive progress is possible.

A Equal Percent Bias Reduction (EPBR)

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 and Thomas 1992)—e.g., distributions such as the normal or t—or if the covariates are mixtures

of proportional ellipsoidally symmetric (DMPES) distributions Rubin and Stuart (2005).[36]

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

$$E(X \mid T = 1) - Z \quad = \gamma \{E(X \mid T = 1) - E(X \mid T = 0)\}$$

for a scalar $0 \leq \gamma \leq 1$. In other words, we say that 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 the unmatched (Rubin 1976a).

Even if the covariates have elliptic distributions, in finite samples they may not. Then Mahalanobis distance may not be optimal because the matrix used to scale the distances, the covariance matrix of X, can be improved upon.

The EPBR property itself is limited and in a given substantive problem it may not be desirable. This can arise if 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 $X_1 > 1$. In such a case, reducing bias in X_1 will be more important than X_2.

[36]Note that DMPES defines a limited set of mixtures. In particular, countably infinite mixtures of ellipsoidal distributions where: (1) all inner products are proportional and (2) where the centers of each constituent ellipsoidal distribution are such that all best linear discriminants between any two components are also proportional.

References

Abadie, Alberto. 2002. "Bootstrap Tests for Distributional Treatment Effect in Instrumental Variable Models." *Journal of the American Statistical Association* 97 (457): 284–292.

Abadie, Alberto and Guido Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74: 235–267.

Barnard, John, Constantine E. Frangakis, Jennifer L. Hill, and Donald B. Rubin. 2003. "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City." *Journal of the American Statistical Association* 98 (462): 299–323.

Bartels, Larry. 1996. "Pooling Disparate Observations." *American Journal of Political Science* 40 (3): 905–942.

Bennett, Andrew. 1998. "Causal Inference in Case Studies: From Mill's Methods to Causal Mechanisms." Paper presented at the annual meeting of the American Political Science Association. Atlanta, GA.

Bonney, Jessica and Brandice Canes-Wrone William Minozzi. 2007. "Issue Accountability and the Mass Public: The Electoral Consequences of Legislative Voting on Crime Policy." Working Paper.

Bowers, Jake and Ben Hansen. 2005. "Attributing Effects to a Get-Out-The-Vote Campaign Using Full Matching and Randomization Inference." Working Paper.

Brady, Henry. 2008. "Models of Causal Inference: Going Beyond the Neyman-Rubin-Holland Theory." In Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, editors, *The Oxford Handbook of Political Methodology* New York: Oxford University Press.

Brady, Henry and Iris Hui. 2006. "Is it Worth Going the Extra Mile to Improve Causal

Inference?".” Paper presented at the 23rd Annual Summer Meeting of the Society of Political Methodology.

Campbell, Donald T. and Julian C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin Company.

Christakis, Nicholas A. and Theodore I. Iwashyna. 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.” *Social Science & Medicine* 57 (3): 465–475.

Cochran, William G. 1953. “Matching in Analytical Studies.” *American Journal of Public Health* 43: 684–691.

Cochran, William G. 1965. “The Planning of Observational Studies of Human Populations (with discussion).” *Journal of the Royal Statistical Society, Series A* 128: 234–255.

Cochran, William G. and Donald B. Rubin. 1973. “Controlling Bias in Observational Studies: A Review.” *Sankhyā,* Ser. A 35: 417–446.

Cohen, Morris and Ernest Nagel. 1934. *An Introduction to Logic and Scientific Method*. Harcourt, Brace and Company.

Cox, David R. 1958. *Planning of Experiments*. New York: Wiley.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2006. “Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand.” Working Paper.

Dawid, A. Phillip. 1979. “Conditional Independence in Statistical Theory.” *Journal of the Royal Statistical Society, Series B* 41 (1): 1–31.

Dawid, A. Phillip. 2000. “Causal Inference without Counterfactuals (with Discussion).” *Journal of the American Statistical Association* 95 (450): 407–424.

Dehejia, Rajeev. 2005. "Practical Propensity Score Matching: A Reply to Smith and Todd." *Journal of Econometrics* 125 (1–2): 355–364.

Dehejia, Rajeev and Sadek Wahba. 1997. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." Rejeev Dehejia, *Econometric Methods for Program Evaluation*. Ph.D. Dissertation, Harvard University, Chapter 1.

Dehejia, Rajeev and Sadek Wahba. 1999. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–1062.

Dehejia, Rajeev H. and Sadek Wahba. 2002. "Propensity Score Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84 (1): 151–161.

Diamond, Alexis and Jasjeet S. Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." See `http://sekhon.berkeley.edu/papers/GenMatch.pdf`.

Diprete, Thomas A. and Henriette Engelhardt. 2004. "Estimating Causal Effects With Matching Methods in the Presence and Absence of Bias Cancellation." *Sociological Methods & Research* 32 (4): 501–528.

Dorn, H. F. 1953. "Philosophy of Inference from Retrospective Studies." *American Journal of Public Health* 43: 692–699.

Eckstein, Harry. 1975. "Case Study and Theory in Political Science." In Fred I. Greenstein and Nelson W. Polsby, editors, *Handbook of Political Science. Vol. 7. Strategies of Inquiry* Reading, MA: Addison-Wesley. pages 79–137.

Fechner, Gustav Theodor. 1966 [1860]. *Elements of psychophysics, Vol 1.*. New York: Rinehart & Winston. Translated by Helmut E. Adler and edited by D.H. Howes and E.G. Boring.

Fisher, Ronald A. 1958a. "Cancer and Smoking." *Nature* 182 (August): 596.

Fisher, Ronald A. 1958b. "Lung Cancer and Cigarettes?" *Nature* 182 (July): 108.

Freedman, David A. 2007a. "On Regression Adjustments in Experiments with Several Treatments." *Annals of Applied Statistics*. In press.

Freedman, David A. 2007b. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics*. In press.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky. 2005. "Water for Life: The Impact of the Privatization of Water Services on Child Mortality." *Journal of Political Economy* 113 (1): 83–120.

George, Alexander L. and Timothy J. McKeown. 1985. "Case Studies and Theories of Organizational Decision-Making." In Robert F. Coulam and Richard A. Smith, editors, *Advances in Information Processing in Organizations* Greenwich, CT: JAI Press. pages 21–58.

Gilligan, Michael J. and Ernest J. Sergenti. 2006. "Evaluating UN Peacekeeping with Matching to Improve Causal Inference." Working Paper.

Gordon, Sandy and Greg Huber. 2007. "The Effect of Electoral Competitiveness on Incumbent Behavior." *Quarterly Journal of Political Science* 2 (2): 107–138.

Green, Donald and Alan Gerber. 2008. "Field Experiments and Natural Experiments." In Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, editors, *The Oxford Handbook of Political Methodology* New York: Oxford University Press.

Hansen, Ben B. 2004. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association* 99: 609–618.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66 (5): 1017–1098.

Herron, Michael C. and Jasjeet S. Sekhon. 2003. "Overvoting and Representation: An examination of overvoted presidential ballots in Broward and Miami-Dade Counties." *Electoral Studies* 22 (1): 21–47.

Herron, Michael C. and Jasjeet S. Sekhon. 2005. "Black Candidates and Black Voters: Assessing the Impact of Candidate Race on Uncounted Vote Rates." *Journal of Politics* 67 (1).

Herron, Michael C. and Jonathan Wand. forthcoming. "Assessing Partisan Bias in Voting Technology: The Case of the 2004 New Hampshire Recount." *Electoral Studies*.

Hill, Bradford. 1961. *Principles of Medical Statistics*. London: The Lancet 7 edition.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–960.

Horvitz, D. G. and D. J. Thompson. 1952. "A Generalization of Sampling without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47: 663–685.

Imai, Kosuke. 2005. "Do Get-Out-The-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review* 99 (2): 283–300.

Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87 (3): 706–710.

Lenz, Gabriel S. and Jonathan McDonald Ladd. 2006. "Exploiting a Rare Shift in Communication Flows: Media Effects in the 1997 British Election." Working Paper.

Lieberson, Stanley. 1991. "Small N's and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases." *Social Forces* 70 (2): 307–320.

Little, Daniel. 1998. *Microfoundations, Method, and Causation*. New Brunswick, NJ: Transaction Publishers.

Mauldon, Jane, Jan Malvin, Jon Stiles, Nancy Nicosia, and Eva Seto. 2000. "Impact of California's Cal-Learn Demonstration Project: Final Report." UC DATA Archive and Technical Assistance.

McKeown, Timothy J. 1999. "Case Studies and the Statistical Worldview: Review of King, Keohane, and Verba's *Designing Social Inquiry: Scientific Inference in Qualitative Research*." *International Organization* 51 (1): 161–190.

Mebane, Walter R. Jr. and Jasjeet S. Sekhon. 1998. "GENetic Optimization Using Derivatives (GENOUD)." Software Package. `http://sekhon.berkeley.edu/rgenoud/`.

Mebane, Walter R. Jr. and Jasjeet S. Sekhon. 2004. "Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data." *American Journal of Political Science* 48 (2): 391–410.

Mill, John Stuart. 1872. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. London: Longmans, Green and Co. 8th edition.

Mitchell, Ann F. S. and Wojtek J. Krzanowski. 1985. "The Mahalanobis Distance and Elliptic Distributions." *Biometrika* 72 (2): 464–467.

Mitchell, Ann F. S. and Wojtek J. Krzanowski. 1989. "Amendments and Corrections: The Mahalanobis Distance and Elliptic Distributions." *Biometrika* 76 (2): 407.

Morgan, Stephen L. and David J. Harding. 2006. "Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice." *Sociological Methods & Research* 35 (1): 3–60.

Neyman, Jerzy. 1923 [1990]. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5 (4): 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed.

Park, Johng Hee. 2006. "Causal Effect of Information on Voting Behavior from a Natural Experiment: An Analysis of Candidate Blacklisting Campaign in 2000 South Korean National Assembly Election." Working paper.

Pledge, Humphry Thomas. 1939. *Science Since 1500: A Short History of Mathematics, Physics, Chemistry [and] Biology*. London: His Majesty's Stationery Office.

Przeworski, A. and H. Teune. 1970. *The Logic of Comparative Social Inquiry*. New York: Wiley.

Raessler, S. and D. B. Rubin. 2005. "Complications when using nonrandomized job training data to draw causal inferences." *Proceedings of the International Statistical Institute*.

Ragin, Charles C. 2000. *Fuzzy-Set Social Science*. Chicago: University of Chicago Press.

Ragin, Charles C., Dirk Berg-Schlosser, and Gisèle de Meur. 1996. "Political Methodology: Qualitative Methods." In Robert E. Goodin and Hans-Dieter Klingemann, editors, *A New Handbook of Political Science* New York: Oxford University Press. pages 749–768.

Robinson, William S. 1951. "The Logical Structure of Analytic Induction." *American Sociological Review* 16 (6): 812–818.

Rosenbaum, Paul R. 1991. "A Characterization of Optimal Designs for Observational Studies." *Journal of the Royal Statistical Society, Series B* 53 (3): 597–610.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer-Verlag 2nd edition.

Rosenbaum, Paul R. 2005. "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies." *The American Statistician* 59: 147–152.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician* 39 (1): 33–38.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.

Rubin, Donald B. 1976a. "Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples." *Biometrics* 32 (1): 109–120.

Rubin, Donald B. 1976b. "Multivariate Matching Methods That are Equal Percent Bias Reducing, II: Maximums on Bias Reduction for Fixed Sample Sizes." *Biometrics* 32 (1): 121–132.

Rubin, Donald B. 1977. "Assignment to a Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2: 1–26.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6 (1): 34–58.

Rubin, Donald B. 1979. "Using Multivariate Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74: 318–328.

Rubin, Donald B. 1980. "Bias Reduction Using Mahalanobis-Metric Matching." *Biometrics* 36 (2): 293–298.

Rubin, Donald B. 1997. "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Annals of Internal Medicine* 127 (8S): 757–763.

Rubin, Donald B. 2001. "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation." *Health Services & Outcomes Research Methodology* 2 (1): 169–188.

Rubin, Donald B. 2006. *Matched Sampling for Causal Effects*. Cambridge, England: Cambridge University Press.

Rubin, Donald B. and Elizabeth A. Stuart. 2005. "Affinely Invariant Matching Methods with Discriminant Mixtures of Proportional Ellipsoidally Symmetric Distributions." Working Paper.

Rubin, Donald B. and Neal Thomas. 1992. "Affinely Invariant Matching Methods with Ellipsoidal Distributions." *Annals of Statistics* 20 (2): 1079–1093.

Salmon, Wesley C. 1989. *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.

Seawright, Jason. 2004. "Qualitative Comparative Analysis vis-a-vis Regression." Paper presented at the 2004 Meetings of the American Political Science Association.

Sekhon, Jasjeet S. 2004a. "Quality Meets Quantity: Case Studies, Conditional Probability and Counterfactuals." *Perspectives on Politics* 2 (2): 281–293.

Sekhon, Jasjeet S. 2004b. "The Varying Role of Voter Information Across Democratic Societies." Working Paper.
URL `http://sekhon.berkeley.edu/papers/SekhonInformation.pdf`

Sekhon, Jasjeet S. 2006. "The Art of Benchmarking: Evaluating the Performance of R on Linux and OS X." *The Political Methodologist* 14 (1): 15–19.

Sekhon, Jasjeet S. 2007. "Matching: Algorithms and Software for Multivariate and Propensity Score Matching with Balance Optimization via Genetic Search." *Journal of Statistical Software*. In press.

Sekhon, Jasjeet Singh and Walter R. Mebane, Jr. 1998. "Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models." *Political Analysis* 7: 189–203.

Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge, UK: Cambridge University Press.

Smith, Herbert L. 1997. "Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology* 27: 305–353.

Smith, Jeffrey and Petra Todd. 2005a. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1–2): 305–353.

Smith, Jeffrey and Petra Todd. 2005b. "Rejoinder." *Journal of Econometrics* 125 (1–2): 365–375.

Smith, Jeffrey A. and Petra E. Todd. 2001. "Reconciling Conflicting Evidence on the Performance of Propensity Score Matching Methods." *AEA Papers and Proceedings* 91 (2): 112–118.

Steering Committee of the Physicians' Health Study Research Group. 1989. "Final report on the aspirin component of the ongoing Physicians' Health Study." *New England Journal of Medicine* 321 (3): 129–135.

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Vandenbroucke, Jan P. 2001. "In Defense of Case Reports and Case Series." *Annals of Internal Medicine* 134 (4): 330–334.

Waldner, David. 2002. "Anti Anti-Determinism: Or What Happens When Schrödinger's Cat and Lorenz's Butterfly Meet Laplace's Demon in the Study of Political and Economic Development." Paper presented at the Annual Meeting of the American Political Science Association, Boston, MA.

Winship, Christopher and Stephen Morgan. 1999. "The estimation of causal effects from observational data." *Annual Review of Sociology* 25: 659–707.