Causal Inference With General Treatment Regimes: Generalizing the Propensity Score

Kosuke IMAI and David A. VAN DYK

In this article we develop the theoretical properties of the propensity function, which is a generalization of the propensity score of Rosenbaum and Rubin. Methods based on the propensity score have long been used for causal inference in observational studies; they are easy to use and can effectively reduce the bias caused by nonrandom treatment assignment. Although treatment regimes need not be binary in practice, the propensity score methods are generally confined to binary treatment scenarios. Two possible exceptions have been suggested for ordinal and categorical treatments. In this article we develop theory and methods that encompass all of these techniques and widen their applicability by allowing for arbitrary treatment regimes. We illustrate our propensity function methods by applying them to two datasets; we estimate the effect of smoking on medical expenditure and the effect of schooling on wages. We also conduct simulation studies to investigate the performance of our methods.

KEY WORDS: Medical expenditure; Nonrandom treatment assignment; Observational studies; Return to schooling; Subclassification; Treatment effect.

1. INTRODUCTION

Establishing the effect of a treatment that is not randomly assigned is a common goal in empirical research. But the lack of random assignment means that groups with different levels of the treatment variable can systematically differ in important ways other than the observed treatment. Because these differences may exhibit complex correlations with the outcome variable, ascertaining the causal effect of the treatment may be difficult. It is in this setting that the propensity score of Rosenbaum and Rubin (1983b) has found wide applicability in empirical research; in particular, the method has rapidly become popular in the social sciences (e.g., Heckman, Ichimura, and Todd 1998; Lechner 1999; Imai 2004).

The propensity score aims to control for differences between the treatment groups when the treatment is binary; it is defined as the conditional probability of assignment to the treatment group given a set of observed pretreatment variables. Under the assumption of strongly ignorable treatment assignment, multivariate adjustment methods based on the propensity score have the desirable property of effectively reducing the bias that frequently arises in observational studies. In fact, there exists empirical evidence that in certain situations the propensity score method produces more reliable estimates of causal effects than other estimation methods (e.g., Dehejia and Wahba 1999; Imai 2004).

The propensity score is called a *balancing score* because, conditional on the propensity score, the binary treatment assignment and the observed covariates are independent (Rosenbaum and Rubin 1983b). If we further assume the conditional independence between treatment assignment and potential outcomes given the observed covariates, then it is possible to obtain unbiased estimates of treatment effects. In practice, matching or subclassification is used to adjust for the *estimated* propensity score, which is ordinarily generated by logistic regression (Rosenbaum and Rubin 1984, 1985). The advantage of using estimated propensity scores in place of true propensity scores has been discussed at length in the literature (e.g., Rosenbaum 1987; Robins, Rotnitzky, and Zhao 1995; Rubin and Thomas 1996; Heckmen et al. 1998; Hirano, Imbens, and Ridder 2003); see also Section 5.3. Indeed, even in randomized experiments where the randomization scheme specifies the true propensity score, adjusting for the estimated propensity score can reduce the variance of the estimated treatment effect. One of the principle advantages of this method is that adjusting for the propensity score amounts to matching or subclassifying on a scalar, which is significantly easier than matching or subclassifying on many covariates.

In this article we extend and generalize the propensity score method so that it can be applied to arbitrary treatment regimes. The original propensity score was developed to estimate the causal effects of a binary treatment; however, in many observational studies, the treatment may not be binary or even categorical. For example, in clinical trials, one may be interested in estimating the dose-response function where the drug dose may take on a continuum of values (e.g., Efron and Feldman 1991). Alternatively, the treatment may be ordinal. In economics, an important quantity of interest is the effect of schooling on wages, where schooling is measured as years of education in school (e.g., Card 1995). The treatment can also consist of multiple factors and their interactions. In political science, one may be interested in the combined effects of different voter mobilization strategies, such as phone calls and door-to-door visits (e.g., Gerber and Green 2000). Treatment can also be measured in terms of frequency and duration, for example, the health effects of smoking. These examples illustrate the need to extend the propensity score, a prominent methodology of causal inference, for application to general treatment regimes.

Two extensions of the propensity score have been developed to handle a univariate categorical or ordinal treatment variable. (We use the term "ordinal variable" to refer to a discrete variable that takes on ordered values, whereas a "categorical variable" is discrete with possibly unordered values.) Imbens (2000) suggested computing a propensity score for each level

Kosuke Imai is Assistant Professor, Department of Politics, Princeton University, Princeton, NJ 08544 (E-mail: *kimai@princeton.edu*). David A. van Dyk is Associate Professor, Department of Statistics, University of California, Irvine, CA 92697-1250 (E-mail: *dvd@ics.uci.edu*). The authors thank Joshua Angrist, Guido Imbens, Elizabeth Johnson, and Scott Zegar for providing the datasets used in this article. They also thank Jim Alt, Samantha Cook, Jennifer Hill, Dan Ho, Gary King, Donald Rubin, Phil Schrodt, Jas Sekhon, and Elizabeth Stuart for helpful discussions. The comments from the associate editor and anonymous referees significantly improved this article. Research support was provided by National Science Foundation grant DMS-01-04129, the U.S. Census Bureau, and the Princeton University Committee on Research in the Humanities and Social Sciences.

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Theory and Methods
DOI 10.1198/016214504000001187

of a categorical treatment variable; that is, he recommends computing the probability of each treatment given the observed covariates. The mean response under each level of the treatment is estimated as the average of the conditional means given the corresponding propensity score. The effect of the treatment can be studied by comparing the mean responses under the various treatment levels. For an ordinal treatment variable, Joffe and Rosenbaum (1999) proposed and Lu, Zanutto, Hornik, and Rosenbaum (2001) applied a method based on a scalar balancing score; matching subjects on this score tends to balance the covariates. Both of these extensions maintain an important advantage of the approach of Rosenbaum and Rubin (1983b); they effectively balance a potentially high-dimensional covariate by adjusting for a scalar propensity score. [Joffe and Rosenbaum (1999) proposed the possibility of adjusting for a low-dimensional linear propensity score in the context of a univariate ordinal treatment. Imbens (2000) also suggested adjusting for several propensity scores, but with the scores adjusted for one at a time.]

In this article we develop methods and theory that encompass the generalized propensity scores of Imbens (2000) and Joffe and Rosenbaum (1999). Our methods can establish causal effects in observational studies where the treatment is categorical, ordinal, continuous, semicontinuous, or even multivariate. Although, our methods are closely related to those of Joffe and Rosenbaum (1999), we emphasize techniques based on subclassification rather than on the matching used by Lu et al. (2001). Finally, we also are able to effectively balance a high-dimensional covariate by adjusting for a low-dimensional (though perhaps not scalar) propensity score.

This article is organized as follows. In Section 2 we describe the propensity function, our generalization of the propensity score. We apply and evaluate our method using a continuous treatment in Section 3 and a bivariate treatment in Section 4. In Section 5 we illustrate how our methods can improve balance in a randomized design. Finally, we give some concluding remarks in Section 6.

2. METHODOLOGY AND THEORY

2.1 Framework for Causal Inference

For a simple random sample of size n, suppose that we observe a $p \times 1$ vector of pretreatment covariates, $\mathbf{X}_i$, the possibly multivariate value of the treatment received, $\mathbf{T}_i^A$, and the value of the outcome variable associated with this treatment, Y_i. Using the Rubin causal model (Holland 1986) as a framework for causal inference, we define a set of potential outcomes, $\mathcal{Y} = \{Y_i(\mathbf{t}^P), \mathbf{t}^P \in \mathcal{T} \text{ for } i = 1, \ldots, n\}$, where $\mathcal{T}$ is a set of potential treatment values and $Y_i(\mathbf{t}^P)$ is a random variable that maps a particular potential treatment, $\mathbf{t}^P$, to a potential outcome. We treat $\mathbf{t}^P$ as an ordinary variable and $\mathbf{T}_i^A$ as a random variable. (In our general notation, we use boldface for the possibly multivariate treatment variables, $\mathbf{T}^A$ and $\mathbf{t}^P$. In particular examples, however, these variables may be univariate, in which case we do not use boldface. Thus, we use T^A and t^P in place of $\mathbf{T}^A$ and $\mathbf{t}^P$ when the treatment is represented via a univariate quantity.)

To evaluate the effect of the treatment, we rely on the standard two assumptions.

Assumption 1: Stable Unit Treatment Value Assumption (Rubin 1980, 1990). The distribution of potential outcomes for one unit is assumed to be independent of potential treatment status of another unit given the observed covariates.

Assumption 1 excludes the possibility of interference between units and, given the observed covariates, allows us to consider the potential outcomes of one unit to be independent of another unit's treatment status. (Thus we suppress the index, i, in the remainder of the article.) Because the treatment assignment mechanism in most observational studies is unknown, the conditional distribution of $\mathbf{T}^A$ given $\mathbf{X}$ needs to be modeled, usually parametrically. Assumption 2 allows us to model $\mathbf{T}^A$ without conditioning on potential outcomes.

Assumption 2: Strong Ignorability of Treatment Assignment (Rosenbaum and Rubin 1983b). The distribution of the actual treatment does not depend on potential outcomes given the observed covariates. Formally, $p\{\mathbf{T}^A|Y(\mathbf{t}^P), \mathbf{X}\} = p(\mathbf{T}^A|\mathbf{X})$ for all $\mathbf{t}^P \in \mathcal{T}$. Furthermore, $0 < p(\mathbf{T}^A \in \mathcal{A}|\mathbf{X})$ for all $\mathbf{X} \in \mathcal{X}$ and measurable sets $\mathcal{A} \subset \mathcal{T}$ with positive measure.

In practice, ignorability is a nontrivial assumption that should be made only with great care; omitting covariates can seriously bias estimates of causal effects (Rosenbaum and Rubin 1983a; Drake 1993); see also Section 5. For clarity, we maintain Assumptions 1 and 2 and discuss generalization of the propensity score method under these assumptions.

When making causal inference, the distribution $p\{Y(\mathbf{t}^P)|\mathbf{X}\}$ as a function of $\mathbf{t}^P$ and for fixed $\mathbf{X}$, or its average over the population, $p\{Y(\mathbf{t}^P)\} = \int p\{Y(\mathbf{t}^P)|\mathbf{X}\}p(\mathbf{X})\,d\mathbf{X}$, is of primary interest. The fundamental difficulty of causal inference in observational studies is that we observe only one of the potential outcomes, $Y(\mathbf{T}^A = \mathbf{t}^P) \in \mathcal{Y}$. Therefore, in practice we must condition on the observed treatment assignment. But because $\mathbf{T}^A$ and $\mathbf{X}$ are not generally independent, basing inference on $p\{Y(\mathbf{t}^P)|\mathbf{T}^A\} = \int p\{Y(\mathbf{t}^P)|\mathbf{T}^A, \mathbf{X}\}p(\mathbf{X}|\mathbf{T}^A)\,d\mathbf{X}$ often leads to bias. The solution lies in conditioning on the observed covariates; by Assumption 2, $p\{Y(\mathbf{t}^P)|\mathbf{T}^A, \mathbf{X}\} \propto p\{Y(\mathbf{t}^P), \mathbf{T}^A|\mathbf{X}\} = p\{Y(\mathbf{t}^P)|\mathbf{X}\}p\{\mathbf{T}^A|\mathbf{X}\} \propto p\{Y(\mathbf{t}^P)|\mathbf{X}\}$. Thus the average causal effect $E\{Y(\mathbf{t}_1^P) - Y(\mathbf{t}_2^P)|\mathbf{X}\} = E\{Y(\mathbf{t}_1^P)|\mathbf{T}^A = \mathbf{t}_1^P, \mathbf{X}\} - E\{Y(\mathbf{t}_2^P)|\mathbf{T}^A = \mathbf{t}_2^P, \mathbf{X}\}$, where $\mathbf{t}_1^P \neq \mathbf{t}_2^P$, and we obtain valid inference conditional on $\mathbf{X}$ even when we condition on the observed treatment assignment.

In principle, we can model $p\{Y(\mathbf{t}^P)|\mathbf{T}^A = \mathbf{t}^P, \mathbf{X}\}$ directly, but experience shows that even with binary treatments, standard model assumptions, such as linearity, do not suffice, and this misspecification can strongly bias causal inference (Drake 1993; Dehejia and Wahba 1999). Various nonparametric techniques exist; matching and subclassification are commonly used. However, as the dimension of $\mathbf{X}$ increases, matching and subclassification become impossible in practice. The propensity score aids analysis in this regard by reducing the dimension of the variable that is conditioned upon to a scalar. In the next section we generalize the propensity score so that it not only is applicable to arbitrary treatment regimes, but also sufficiently reduces the dimension of $\mathbf{X}$ to allow for efficient matching or subclassification.

2.2 The Propensity Function

We define the propensity function as the conditional probability of the actual, perhaps multivariate, treatment given the observed covariates, that is, $p_{\boldsymbol{\psi}}(\mathbf{T}^A|\mathbf{X})$, where $\boldsymbol{\psi}$ parameterizes this distribution. When T^A is binary the propensity function is determined by the propensity score, $p_{\boldsymbol{\psi}}(T^A|\mathbf{X})|_{T^A=1}$, where T^A is an indicator variable for the treatment. More generally, the parametric model defines the propensity function, $e_{\boldsymbol{\psi}}(\cdot|\mathbf{X}) = p_{\boldsymbol{\psi}}(\cdot|\mathbf{X})$. When the propensity function is unknown, misspecification of the model is possible, which may bias causal inference. Thus care must be taken both to identify as many covariates as possible and to check for model misspecification (Drake 1993); see also Section 5.

To simplify the representation of the propensity function and to facilitate subclassification and matching, we make the following assumption regarding its parameterization.

Assumption 3: Uniquely Parameterized Propensity Function. For every $\mathbf{X} \in \mathcal{X}$, there exists a unique finite-dimensional parameter, $\boldsymbol{\theta} \in \Theta$, such that $e_{\boldsymbol{\psi}}(\cdot|\mathbf{X})$ depends on $\mathbf{X}$ only through $\boldsymbol{\theta}_{\boldsymbol{\psi}}(\mathbf{X})$. That is, $\boldsymbol{\theta}$ uniquely represents $e\{\cdot|\boldsymbol{\theta}_{\boldsymbol{\psi}}(\mathbf{X})\}$, which we may therefore write as $e(\cdot|\boldsymbol{\theta})$.

Under this assumption, the propensity function is characterized by the parameter θ, which is typically of much lower dimension than $\mathbf{X}$. In some cases, $\boldsymbol{\theta}$ is univariate, in which case we write θ in place of $\boldsymbol{\theta}$. To illustrate Assumption 3 and methods based on the propensity function, we consider three simple examples.

Example With a Continuous Treatment. Suppose that we model the distribution of the treatment given a $(p \times 1)$ vector of covariates, $\mathbf{X}$, as $T^A|\mathbf{X} \sim \mathrm{N}(\mathbf{X}^\top\boldsymbol{\beta}, \sigma^2)$ where σ^2 is a scalar and $\boldsymbol{\beta}$ is a $(p \times 1)$ vector. The propensity function, $e\{\cdot|\theta_{\boldsymbol{\psi}}(\mathbf{X})\}$, is the Gaussian density function, $\boldsymbol{\psi} = (\boldsymbol{\beta}, \sigma^2)$, and $\theta_{\boldsymbol{\psi}}(\mathbf{X}) = \mathbf{X}^\top\boldsymbol{\beta}$. Given $\boldsymbol{\psi}$, the propensity function is characterized by the scalar, θ. Hence matching or subclassifying on the propensity function can be easily accomplished by matching or subclassifying on θ or any one-to-one function of θ, regardless of the dimension of $\mathbf{X}$. As a generalization, we can allow σ^2 to be some function of $\mathbf{X}$, but in this case $\theta_{\boldsymbol{\psi}}(\mathbf{X})$ would usually be multivariate.

Example With a Categorical Treatment. The propensity function also encompasses the propensity scores suggested by Imbens (2000) for a categorical treatment. Suppose that $\mathcal{T} = \{1, \ldots, t_{\max}\}$ and we model $p_{\boldsymbol{\psi}}(T^A|\mathbf{X})$ as a multinomial distribution with probability vector $\boldsymbol{\pi}(\mathbf{X}) = \{\pi_1(\mathbf{X}), \ldots, \pi_{\max}(\mathbf{X})\}$. If for each $\mathbf{X}$, $\boldsymbol{\pi}(\mathbf{X})$ is a probability vector without any additional constraints, then $\boldsymbol{\theta}_{\boldsymbol{\psi}}(\mathbf{X}) = \boldsymbol{\pi}(\mathbf{X})$ is a $t_{\max}$-dimensional parameter that corresponds to the set of $t_{\max}$ propensity scores proposed by Imbens (2000). We might use nested logistic regression (as suggested in Imbens 2000) or a multinomial probit model (e.g., Imai and van Dyk 2004) to model the dependence of $\boldsymbol{\pi}(\mathbf{X})$ on $\mathbf{X}$; in either case, $\boldsymbol{\psi}$ represents the regression coefficients.

Example With an Ordinal Treatment. The propensity score suggested by Joffe and Rosenbaum (1999) for an ordinal treatment is also a special case of the propensity function. We can use the same setup as in the example with a categorical treatment, except that we model $\boldsymbol{\pi}(\mathbf{X})$ using an ordinal logistic model (McCullagh and Nelder 1989). In this case $\boldsymbol{\pi}(\mathbf{X})$ is determined by the scalar $\mathbf{X}^\top\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a $(p \times 1)$ parameter vector; in the general framework $\boldsymbol{\psi} = \boldsymbol{\beta}$ and $\theta_{\boldsymbol{\psi}}(\mathbf{X}) = \mathbf{X}^\top\boldsymbol{\beta}$. Lu et al. (2001) mentioned the possibility of using Gaussian linear regression to model the assignment mechanism for an ordinal treatment, but then constant residual variance must be assumed. This constraint is not necessary under our general framework, but allowing for nonconstant variance generally increases the dimension of $\boldsymbol{\theta}_{\boldsymbol{\psi}}(\mathbf{X})$.

2.3 Large-Sample Theory

Under the analytical framework and assumptions given in Sections 2.1 and 2.2, we derive theoretical results that closely follow and extend those in of Rosenbaum and Rubin (1983b); these results are verified in Appendix A. Throughout we assume that the propensity function including the parameters, $\boldsymbol{\psi}$, is known. Result 1 states that the propensity function is a balancing score even with a nonbinary treatment; that is, given the propensity function, the conditional distribution of the actual treatment does not depend on the covariates.

Result 1: Propensity Function as a Balancing Score.

$$p(\mathbf{T}^A|\mathbf{X}) = p\{\mathbf{T}^A|\mathbf{X}, e(\cdot|\mathbf{X})\} = p\{\mathbf{T}^A|e(\cdot|\mathbf{X})\}.$$

In practice, Result 1 can be checked, for example, by examining the t-statistics for the coefficient of T^A in models that predict each covariate while controlling for the estimated propensity function. We use this diagnostic in Sections 3–5. We emphasize, however, that diagnostics based solely on the t-statistics of a linear model do not detect all deviations from independence, and in some cases more sophisticated diagnostics (for example, based on nonlinear transformations of the covariates) may be required; see Section 3.2 and Appendix B.

We can now establish the key result, which states that the potential outcomes and the actual treatment assignment are conditionally independent given the propensity function.

Result 2: Strong Ignorability of Treatment Assignment Given the Propensity Function. $p\{Y(\mathbf{t}^P)|\mathbf{T}^A, e(\cdot|\mathbf{X})\} = p\{Y(\mathbf{t}^P)|e(\cdot|\mathbf{X})\}$ for any $\mathbf{t}^P \in \mathcal{T}$.

We can average $p\{Y(\mathbf{t}^P)|e(\cdot|\mathbf{X})\}$ over the distribution of the propensity function to obtain the distribution of primary interest, $p\{Y(\mathbf{t}^P)\}$ as a function of $\mathbf{t}^P$. According to Result 2,

$$p\{Y(\mathbf{t}^P)\} = \int p\{Y(\mathbf{t}^P)|\mathbf{T}^A = \mathbf{t}^P, \boldsymbol{\theta}\}p(\boldsymbol{\theta})\,d\boldsymbol{\theta}, \tag{1}$$

where $\boldsymbol{\theta} = \boldsymbol{\theta}_{\boldsymbol{\psi}}(\mathbf{X})$ uniquely indexes the propensity function.

2.4 From Theory to Practice

Subclassification. We can approximate the integration in (1) by subclassifying similar values of $\boldsymbol{\theta}$. In particular, we first model $p_{\boldsymbol{\psi}}(\mathbf{T}^A|\mathbf{X})$ and compute the estimate $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$, perhaps by maximum likelihood (ML). We then compute $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\hat{\boldsymbol{\psi}}}(\mathbf{X})$ for each observation and subclassify observations with the same or similar values of $\hat{\theta}$ into a moderate number, say J, of subclasses of roughly equal size. Within each subclass, we

parametrically model $p_{\boldsymbol{\phi}}\{Y(\mathbf{t}^P)|\mathbf{T}^A = \mathbf{t}^P\}$, where $\boldsymbol{\phi}$ is an unknown parameter. The distributions of the potential outcomes can be computed as a weighted averages of the within-subclass distributions with weights equal to the relative size of the subclasses. Formally, we approximate (1) with

$$p\{Y(\mathbf{t}^P)\} = \int p\{Y(\mathbf{t}^P)|\mathbf{T}^A = \mathbf{t}^P, \boldsymbol{\theta}\} p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$
$$\approx \sum_{j=1}^{J} p_{\hat{\boldsymbol{\phi}}_j}\{Y(\mathbf{t}^P)|\mathbf{T}^A = \mathbf{t}^P\}\, W_j, \tag{2}$$

where $\hat{\boldsymbol{\phi}}_j$ is estimate of $\boldsymbol{\phi}$ in subclass j and W_j is the relative weight of subclass j.

Equation (2) shows how we can approximate the marginal distributions of the potential outcomes. Although these distributions are sometimes appropriate in practice (e.g., Imbens and Rubin 1997), more often they are summarized by the relevant causal effect. This causal effect is generally a function of $\boldsymbol{\phi}$, for instance, the regression coefficient of $Y(t^P)$ on t^P. In practice, additional adjustment within each subclass is desirable to further reduce bias. For example, some authors suggest adjusting for the covariates in the within-subclass model (e.g., Robins and Rotnitzky 2001). We believe that this is generally a useful strategy for accounting for the within-subclass variability of $\boldsymbol{\theta}$, and thus we include available covariates when fitting the within-subclass models in our simulations and examples in Sections 3–5. In particular, we compute

$$\hat{\boldsymbol{\phi}} \approx \sum_{j=1}^{J} \hat{\phi}_j\{Y(\mathbf{t}^p)|\mathbf{T}^A = \mathbf{t}^P, \mathbf{X}\} W_j, \tag{3}$$

where $\hat{\boldsymbol{\phi}}_j\{Y(\mathbf{t}^p)|\mathbf{T}^A = \mathbf{t}^P, \mathbf{X}\}$ is the estimated model parameter in subclass j and the estimated causal effect is some function of $\hat{\boldsymbol{\phi}}$. If W_j is known and the estimate of the causal effect is unbiased within each subclass, then this procedure results in an unbiased estimate of the causal effect. In practice, we estimate W_j by the relative proportion of the observations that fall into subclass j. Because results may be sensitive to the number of subclasses and the choice of subclassification on $\hat{\boldsymbol{\theta}}$, we suggest conducting a sensitivity analysis, repeating the analysis with different subclassification schemes.

Smooth Coefficient Models. Under subclassification, we compute $\hat{\boldsymbol{\phi}}_j$ separately in a number of subclasses. This strategy is robust because it is inherently nonparametric; it avoids modeling the dependence of $p\{Y(\mathbf{t}^P)|\mathbf{T}^A = \mathbf{t}^P, \boldsymbol{\theta}\}$ on $\boldsymbol{\theta}$. Of course, other nonparametric strategies are also possible. One particularly promising possibility was suggested by a referee. Rather than fixing $\boldsymbol{\phi}$ in each of several subclasses, we can allow $\boldsymbol{\phi}$ to vary smoothly as a function of $\boldsymbol{\theta}$; that is, by computing $\boldsymbol{\phi}(\boldsymbol{\theta})$ using a flexible model, such as *penalized regression splines* (e.g., Wood 2003). In the context of linear regression, we can allow the regression coefficients to vary with $\boldsymbol{\theta}$, in what is known as a smooth coefficient model (DiNardo and Tobias 2001; Li, Huang, Li, and Fu 2002; Yatchew 1998). We illustrate this strategy with a continuous treatment in Section 3.3 and with a bivariate treatment in Section 4.3.

Known Propensity Functions. Even if the true propensity function is known, adjusting for the estimated propensity function can be advantageous. Indeed, there is a large literature on the advantage of adjusting for the estimated propensity score rather than the true propensity score in both observational studies and randomized experiments (e.g., Rosenbaum 1987; Rubin and Thomas 1992, 1996; Hill, Rubin, and Thomas 1999). The advantage of the estimated propensity score can be understood by identifying two types of errors that can occur when estimating treatment effects. First, there may be a systematic relationship between the distribution of the covariates and the treatment. Second, there may be random differences in the distribution of the covariates as a function of the treatment in the *observed sample*. Such random differences would average to 0 over repeated sampling but are nonetheless present in any particular sample. Adjusting for either the true or the estimated propensity score accounts for systematic relationships between the covariates and the treatment, but only adjusting for the estimated propensity score can account for sample-specific random differences (Rubin and Thomas 1992; Hill et al. 1999). It is to adjust for such sample-specific differences that covariate adjustment methods, such as ANCOVA, are used to analyze randomized experiments. But adjusting for the propensity score rather than directly adjusting for the covariates has the same advantage in randomized experiments as it has in observational studies: dimension reduction allows for nonparametric adjustment methods such as matching and subclassification. Thus, even in randomized experiments, adjusting for the estimated propensity function can improve estimated treatment effects. We illustrate an application with randomized treatment assignment in Section 5.3.

3. EFFECTS OF SMOKING USING A CONTINUOUS TREATMENT

3.1 Background, Data, and Previous Studies

As a first applied example, we estimate the average effect of smoking on annual medical expenditures. Associated with lawsuits against the tobacco industry, many recent studies have estimated the effects of smoking on health and medical costs (see, e.g., Rubin 2000, 2001; Zeger, Wyant, Miller, and Samet 2000; references therein). The lack of experimental data led many researchers to use propensity scores. Because this method is confined to a binary treatment, the focus has been on the comparison of smokers and nonsmokers without distinguishing among smokers based on how much they smoke (e.g., Larsen 1999; Rubin 2001). In contrast, our proposed method can estimate the causal effects of the frequency and duration of smoking.

We use the data that Johnson, Dominici, Griswold, and Zeger (2003) extracted from the 1987 National Medical Expenditure Survey (NMES). The advantages of the NMES are that it includes detailed information about frequency and duration of smoking, and that 1987 medical costs are verified by multiple interviews and additional data from clinicians and hospitals. Our analysis includes the following subject-level covariates: age at the times of the survey (19–94), age when the individual started smoking, gender (male, female), race (white, black, other), marriage status (married, widowed, divorced, separated,

never married), education level (college graduate, some college, high school graduate, other), census region (Northeast, Midwest, South, West), poverty status (poor, near poor, low income, middle income, high income), and seat belt usage (rarely, sometimes, always/almost always).

As in the original study reported by Johnson et al. (2003), we conduct a complete-case analysis by discarding all individuals with nonresponse. Johnson et al. (2003) noted that better accounting for the missing data using multiple imputation did not significantly affect their results. In general, complete-case analysis involving the propensity score produces biased causal inference unless the data are missing completely at random (D'Agostino and Rubin 2000). Nonetheless, because our purpose is to illustrate the use of the propensity function, we focus on the complete-case analysis, yielding a sample of 9,073 smokers.

The original study did not directly estimate the effects of smoking on medical expenditure. Rather, the authors first estimated the effects of smoking on certain diseases and then examined how much those diseases increased medical costs. In contrast, we directly estimate the effects of smoking on medical expenditures. We focus on smokers and explore the effects of two types of treatment variables. First, we use a measure of the cumulative exposure to smoking to differentiate among smokers according to how much they smoke. Johnson et al. (2003) proposed a measure of cumulative exposure to smoking that combines self-reported information about frequency and duration of smoking. This variable, called *packyear*, is defined as

$$packyear = \frac{\text{number of cigarettes per day}}{20} \times \text{number of years smoked.} \quad (4)$$

In this section we use $\log(packyear)$ as the treatment variable. In Section 4 we consider a bivariate treatment that assess the effects of both the duration and the frequency of smoking.

3.2 A Simulation Study

Before we analyze the actual data, we conduct simulation studies to illustrate how subclassifying on the estimated propensity function can improve the statistical properties of estimated causal effects. In this simulation we use the covariates and treatment variables collected in the NMES, as described earlier. To assess our methods, however, we generate the outcome variable using various known functions of the covariates and treatment variables. In particular, we follow others (e.g., Rubin 1973, 1979; Rubin and Thomas 2000) and use an exponential function to create models with varying degrees of nonlinearity and nonadditivity. Specifically, we closely follow the simulation studies described by Rubin and Thomas (2000) by constructing an additive model of the form $Y_i = \alpha_i T_i^A + c_1(\boldsymbol{\lambda}) \sum_{p=1}^{P} \lambda_p \exp(\kappa_p X_{ip})$ and a multiplicative model of the form, $Y_i = \alpha_i T_i^A + c_2(\boldsymbol{\lambda}) \exp(\sum_{p=1}^{P} \lambda_p X_{ip})$, where Y_i, α_i, T_i^A, and X_{ip} are the response variable, the treatment effect, the assigned treatment, and the the pth covariate for unit i.

In both models the coefficient vector for covariates is represented by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_P)$, the constants $c_1(\boldsymbol{\lambda})$ and $c_2(\boldsymbol{\lambda})$ determine the relative influence of the treatment and the covariates on the response, and each component of $\boldsymbol{\kappa} = (\kappa, \ldots, \kappa_P)$ is either $+1$ or -1. In each replication, every component of $\boldsymbol{\lambda}$ is drawn independently from a Gaussian distribution with mean 1 in the additive model and mean 0 in the multiplicative model; the variance of the Gaussian distribution is specified so as to achieve the desired degree of nonlinearity. Finally, each component of $\boldsymbol{\kappa}$ is set to $+1$ or -1 independently and with equal probability.

We measure deviation from linearity using the squared correlation coefficient, R^2, calculated by regressing each of the nonlinear functions on the set of covariates. We examine three degrees of nonlinearity: highly linear ($R^2 \approx .95$), moderately linear ($R^2 \approx .85$), and moderately nonlinear ($R^2 \approx .75$). Rubin and Thomas (2000, p. 585) noted that detecting the degree of nonlinearity corresponding to $R^2 \approx .85$ is difficult in realistic multivariate settings. The constants $c_1(\boldsymbol{\lambda})$ and $c_2(\boldsymbol{\lambda})$ are set such that the variance of the simulated outcome variable is roughly equal to the variance of the observed outcome variable in the dataset, the log transformation of positive medical expenditure. This implies that the contribution of the nonlinear functions to the overall variance of the simulated outcome variable is about 40%.

Finally, we conduct the simulation under two scenarios, constant treatment effect and variable treatment effect. Under the constant treatment effect scenario, we set α_i to be the same for all individuals using the estimate from the direct Gaussian linear regression model shown in Table 2 (see Sec. 3.3). In the variable treatment effect scenario, we let α_i vary as a function of the the covariates. In particular, we set $\alpha_i = (X_i - \bar{X})^2/1{,}000$, where X_i is the age at which smoking began and $\bar{X}$ is the sample mean of this variable; the scaling makes the average treatment effect close to the value used in the constant treatment scenario. We obtain 1,000 sets of simulated responses for each of the 12 nonlinear models that correspond to the rows of Table 1 (see below in this section).

First, we apply the propensity function method using Gaussian linear regression to model the distribution of the treatment variable given all of the covariates, $\mathbf{X}$. Because only the response variable changes among the simulated datasets, this distribution is the same across the simulations and in the actual data analysis described in Section 3.3. The estimated propensity function is uniquely determined by the linear predictor, $\hat{\theta} = \mathbf{X}^\top \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the ML estimate of the regression coefficients. To evaluate the balance of the covariates, we regress each covariate on the treatment variable, $T^A = \log(packyear)$, using logistic linear and Gaussian linear regression for indicator and continuous covariates. [We use the log transformation of continuous covariates because $\log(packyear)$ and each covariate is necessarily uncorrelated given $\hat{\theta}$; see App. B.] Figure 1(a) shows a standard normal quantile plot of the t-statistics (df = 9,071) for the coefficient of the treatment variable in each regression. The lack of balance is evident in the magnitude of the t-statistics; the treatment variable is highly correlated with many of the covariates. Figure 1(b) is identical to the 1(a) except that we control for $\hat{\theta}$ in each regression. The figure shows the substantial reduction in the t-statistics obtained by conditioning on the estimated propensity function, indicating that the covariate balance is significantly improved. The quantile plots in Figure 1 are constructed including the square of the two age covariates. Including these variables improves the balance; if they are not included, then the t-statistic for $\log(packyear)$ as a

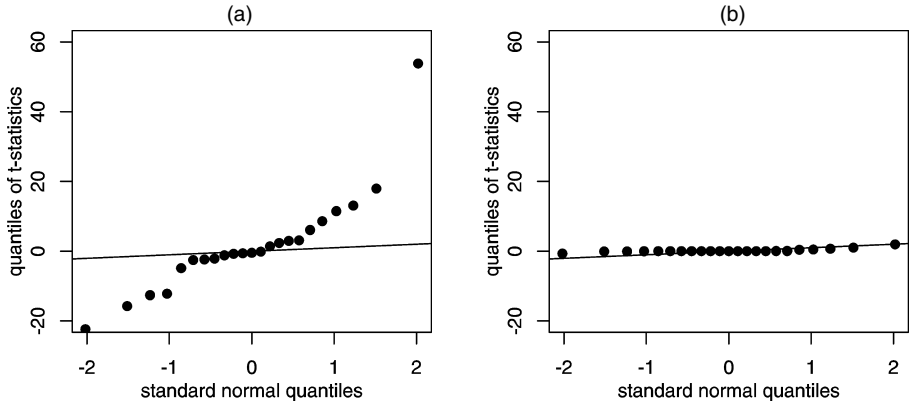

Figure 1. Standard Normal Quantile Plots of t-Statistics for the Coefficient of log(packyear) in the Models Predicting Each Covariate, (a) Without Controlling for $\hat{\theta}$ and (b) Controlling for θ, the Linear Predictor of the Estimated Propensity Function.

predictor of the log of subject age is 6.33 even after controlling for $\hat{\theta}$; including the square terms reduces this t-statistic to .40. This is an example of a check of the propensity function as suggested in Sections 2.2 and 2.3.

Using the resulting estimated propensity score, we construct subclasses of roughly equal size. As suggested in Section 2.4, within each subclass we fit the same Gaussian linear regression using all covariates to estimate the treatment effect. We then obtain the overall average treatment effect by computing the weighted average of the within-subclass treatment effects. To assess the sensitivity of subclassification schemes, we use 3, 5, and 10 subclasses.

Table 1 compares the performance of subclassification on the estimated propensity function with the direct Gaussian linear regression of Y on T^A and $\mathbf{X}$. In particular, we compute the percent reduction in bias and mean squared error (MSE) under 36 different settings (12 models times three subclassification schemes). Overall, subclassification on the propensity function significantly improves the regression estimate; it reduces bias by 16–95%. It is not surprising that in cases when the assumptions of the direct regression model are appropriate (i.e., additive models with a constant treatment effect), this model results in more efficient estimates than subclassifying on the propensity function, which makes fewer model assumptions; robust methods are generally less efficient. But even in these cases, the simulation indicates that subclassification reduces bias. We emphasize that in practice, we would expect the assumptions of the linear model to be violated. For example, in our data examples in Sections 3.3 and 4.3, the treatment effect does not appear to be constant. The simulation illustrates that subclassification on the propensity function can successfully reduce bias and improve efficiency of a parametric model. It also indicates that the propensity function method can be more robust to model misspecification than direct linear regression. In this simulation, the advantage of the propensity score method deteriorates markedly if the covariates are not included in the within-subclass regressions; thus, we recommend including the covariates.

Table 1. Performance of Subclassification on the Estimated Propensity Function Relative to Direct Linear Regression

	3 subclasses		*5 subclasses*		*10 subclasses*	
	Bias	*MSE*	*Bias*	*MSE*	*Bias*	*MSE*
Constant treatment effect						
Additive models						
Highly linear	67	20	82	24	94	18
Moderately linear	67	1	82	5	93	−2
Moderately nonlinear	69	−6	83	−4	95	−12
Multiplicative models						
Highly linear	52	26	69	30	75	24
Moderately linear	80	26	91	21	84	23
Moderately nonlinear	69	38	79	37	88	43
Variable treatment effect						
Additive models						
Highly linear	16	26	29	45	16	26
Moderately linear	17	24	30	40	17	23
Moderately nonlinear	16	19	28	35	16	18
Multiplicative models						
Highly linear	22	38	36	55	26	41
Moderately linear	19	40	31	51	20	41
Moderately nonlinear	22	31	34	44	24	33

NOTE: This table shows the percent reduction in bias and MSE due to subclassification relative to direct linear regression. The covariates and treatment variable are from the dataset of Johnson et al. (2003), and the results are based on 1,000 replications of the response variable simulated under each of the 12 models.

3.3 Data Analysis

We now turn to the observed response variable, self-reported medical expenditure, denoted by Y. We use the same propensity function as that described in Section 3.2 and model Y within each of 10 subclasses using the two-part model of Duan, Manning, Morris, and Newhouse (1983) for semicontinuous variables. In particular, within each subclass we first model the probability of spending some money on medical care, $\Pr(Y > 0|T^A, \mathbf{X})$, given the treatment variable, $T^A = \log(\textit{packyear})$, and all covariates, $\mathbf{X}$, using logistic regression. We then model the conditional distribution of $\log(Y)$, given T^A and $\mathbf{X}$, for those individuals who reported positive medical expenditure, $p(\log(Y)|Y > 0, T^A, \mathbf{X})$, using Gaussian linear regression (see also Olsen and Schafer 2001; Javaras and van Dyk 2003).

Using this two-part model, we estimate the effects of smoking on medical costs within each of the 10 subclasses. Finally, we compute the weighted average of the 10 within-subclass estimates to obtain the average causal effect; in each within-subclass analysis, we use the sampling weights provided in the dataset. Using three subclasses produces similar results, as given in Table 2. We also fit a smooth coefficient model by letting the causal effect as well as an intercept vary smoothly as a function of $\hat{\theta}$. In parallel with the aforementioned two-part model, we use the generalized additive model (Hastie and Tibshirani 1990) with the binomial family and logistic link to model the probability of positive medical costs, and use the Gaussian family and identity link to model the conditional distribution of $\log(Y)$. We fit these models using all of the covariates with the R package *mgcv* developed by Simon Wood; excluding the covariates in this model has little effect on the fitted causal effects.

Table 2 presents the results from the methods based on the propensity function as well as the results of the standard complete-case linear and logistic regressions, both of which include all covariates. All methods agree that cumulative exposure to smoking, as measured by the *packyear* variable, has little effect on the probability of spending some money on medical care in 1987. In contrast, we find that smoking appears to increase medical expenditure among those who reported positive medical cost. (As pointed out by a referee, this ignores the fact that smoking can be fatal and potentially reduce medical expenditure.) Moreover, the two methods based on the propensity function yield a greater effect of smoking on medical expenditure than the standard linear regression analysis. In particular, if *packyear* were to double, then we would expect annual medical expenditure to increase by a factor of about 1.04.

Table 2. Estimated Average Causal Effect of Increased Smoking on Medical Costs

		Propensity score methods		
	Direct models	*3 subclasses*	*10 subclasses*	*Smooth coefficient model*
Logistic regression model				
Coefficient for T^A	−.085	−.082	−.079	−.070
Standard error	3.075	2.996	3.126	3.260
Gaussian regression model				
Average causal effect	.029	.044	.048	.050
Standard error	.017	.017	.018	.017

NOTE: The logistic regression model presents the coefficient of the treatment variable for predicting positive medical costs. The Gaussian regression model presents the estimated average causal effects of log(*packyear*) on log(*medical expenditure*) for individuals with positive medical costs.

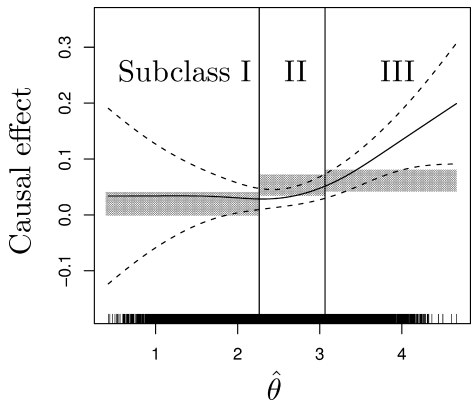

Figure 2. Estimated Causal Effect From the Smooth Coefficient Model. The solid curve represents the causal effect as a function of $\hat{\theta}$ and is based on the estimated coefficient of log(packyears) from the (Gaussian) smooth coefficient model presented in Table 2. The dotted lines show two standard errors above and below the estimate. The vertical lines represent division into three subclasses of equal size. The observed value of $\hat{\theta}_i$ are indicated by short bars on the horizontal axis. The gray bands correspond to two standard errors above and below the within-subclass estimates based on the within-subclass Gaussian linear regressions.

Figure 2 illustrates an advantage of using the propensity score methods in this example. The figure plots the estimated causal effect from the smooth coefficient model as a function of the estimated propensity function. The constant treatment effect assumption of the standard regression models is not appropriate here; the constant treatment effect model is rejected with approximate $p < .002$ under the smooth coefficient model. (This p value was computed with the *mgcv* package in *R*.) The two propensity score methods presented in this section relax this assumption. Subclassification enables us to estimate the causal effect separately within each subclass, whereas the smooth coefficient model allows the causal effect to vary smoothly as a function of $\hat{\theta}$. In this case, age is highly correlated with the assigned treatment. Thus, roughly speaking, Figure 2 shows that as age increases, the effect of log(*packyear*) on medical expenses also increases.

4. EFFECTS OF SMOKING USING A BIVARIATE TREATMENT

4.1 The Bivariate Treatment

Instead of combining frequency and duration into a single measure, we can conduct an analysis with a bivariate treatment composed of the duration of smoking (the log number of smoking months) and the frequency of smoking (the log number of cigarettes per day).

4.2 A Simulation Study

Before analyzing the data using the bivariate treatment, we conduct a simulation study using the same setup as in Section 3.2, except that the single treatment variable is replaced by the sum of two treatment variables. In particular, we construct an additive model of the form $Y_i = \alpha_{i1} T_{i1}^A + \alpha_{i2} T_{i2}^A +$

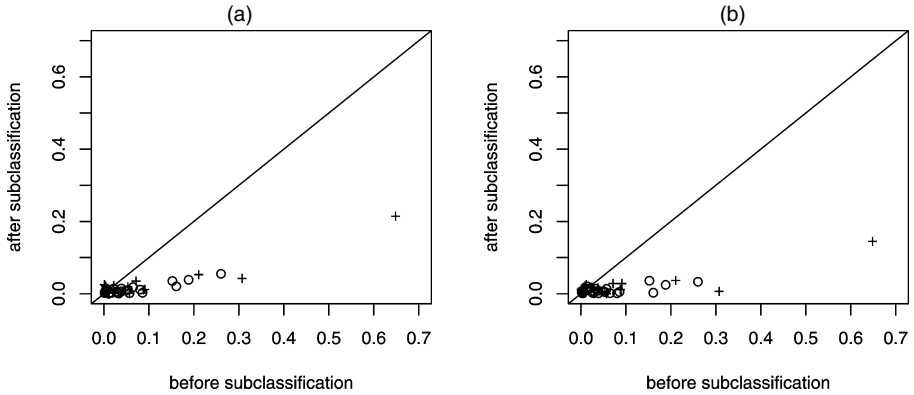

Figure 3. Reduction in Correlations Between the Two Treatment Variables and the Covariates for (a) 3×3 and (b) 4 × 4 Subclasses. The panels plot the absolute value of the correlations between the each of the two treatment variables and each of the covariates (horizontal axis) against the average of the absolute value of the within-subclass correlations (vertical axis). The circles indicate the correlations between the duration treatment variable and the covariates, whereas the crosses represent the correlations between the frequency treatment variable and the covariates.

$c_1(\boldsymbol{\lambda}) \sum_{p=1}^{P} \lambda_p \exp(\kappa_p X_{ip})$ and a multiplicative model of the form $Y_i = \alpha_{i1} T_{i1}^A + \alpha_{i2} T_{i2}^A + c_2(\boldsymbol{\lambda}) \exp(\sum_{p=1}^{P} \lambda_p X_{ip})$, where T_{i1}^A and T_{i2}^A represent the the duration and the frequency of smoking for individual i and α_{i1} and α_{i2} are the corresponding treatment effects. As in Section 3.2, we simulate 1,000 sets of response variables for each of the 12 nonlinear models. To construct the variable treatment effect models, we set α_{i1} equal to the variable treatment in Section 3.2 and construct α_{i2} in the same manner except using the current age covariate.

We first estimate two propensity functions, one function for the frequency of smoking and one for the duration of smoking. We model the propensity functions using two independent Gaussian linear regression models and fit the models via ML and the propensity functions summarized by $\hat{\theta}_1 = \mathbf{X}^\top \hat{\boldsymbol{\beta}}_1$ and $\hat{\theta}_2 = \mathbf{X}^\top \hat{\boldsymbol{\beta}}_2$, with $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ representing the ML estimate of the covariates for the two models. In addition to the set of covariates, we include the square terms for the two age variables in both models to improve the balance given the two linear predictors. Figure 3 shows the significant reduction in correlations between the treatment variables and each of the covariates. In particular, after subclassification on the propensity function, the absolute magnitude of the mean within-subclass correlation is less than .1 for all variables except one of the age variables, whose correlation is reduced by 2/3.

We subclassify the data into several subclasses based on $\hat{\theta}_1$ and $\hat{\theta}_2$. Each subclass contains units with a specific range of both $\hat{\theta}_1$ and $\hat{\theta}_2$. As Figure 4 illustrates, in the 3×3 table of subclasses the first subclass contains units with $\hat{\theta}_1$ and $\hat{\theta}_2$ lower than their 33rd percentile, and the last subclass contains units with both quantities above their 67th percentile. (In some cases, classification schemes that are more complex than a simple grid may be required.) Next, we estimate the average causal effects within each subclass using Gaussian linear regression. Namely,

Propensity function for frequency	Propensity function for duration: Lower third	Middle third	Upper third
Upper third	**Subclass I** duration: .317 (.221) frequency: −.223 (.143) $n = 324$	**Subclass II** duration: .075 (.092) frequency: .125 (.075) $n = 1{,}160$	**Subclass III** duration: .016 (.078) frequency: .093 (.067) $n = 1{,}542$
Middle third	**Subclass IV** duration: .020 (.105) frequency: .009 (.075) $n = 1{,}162$	**Subclass V** duration: −.011 (.092) frequency: .123 (.076) $n = 910$	**Subclass VI** duration: −.182 (.100) frequency: .208 (.080) $n = 952$
Lower third	**Subclass VII** duration: −.079 (.099) frequency: .105 (.058) $n = 1{,}538$	**Subclass VIII** duration: −.178 (.096) frequency: .016 (.072) $n = 954$	**Subclass XI** duration: .018 (.138) frequency: .026 (.106) $n = 532$

Figure 4. Within-Subclass Estimates of the Causal Effects of Smoking on Medical Expenditure. Each cell of the 3 × 3 table represents a subclass within which units have a particular range of the propensity functions for the two treatments. The vertical and horizontal lines that form the subclasses are the 33rd and 67th percentiles of the two propensity functions. The figures within each cell represent the estimated coefficients from the within-subclass Gaussian linear regression and the number of within-subclass observations; standard errors are in parentheses, and n represents the subclass sample sizes.

Table 3. Performance of Subclassification on the Estimated Propensity Function Compared With Linear Regression

	2 × 2 subclasses		3 × 3 subclasses		4 × 4 subclasses	
	Duration	*Frequency*	*Duration*	*Frequency*	*Duration*	*Frequency*
Constant treatment effect						
Additive models						
Highly linear	72	53	87	96	92	92
Moderately linear	62	53	74	97	82	88
Moderately nonlinear	66	24	81	43	87	30
Multiplicative models						
Highly linear	67	31	80	24	86	3
Moderately linear	86	18	77	29	72	54
Moderately nonlinear	82	1	98	38	75	21
Variable treatment effect						
Additive models						
Highly linear	77	22	94	69	99	74
Moderately linear	77	22	93	68	98	74
Moderately nonlinear	77	21	94	68	99	74
Multiplicative models						
Highly linear	77	21	93	68	99	74
Moderately linear	77	22	94	67	99	72
Moderately nonlinear	78	21	94	69	99	74

NOTE: The figures show the percent reduction in bias due to subclassification on the estimated propensity function in comparison with linear regression. The covariates and treatment variable are from the dataset of Johnson et al. (2003), and the results are based on 1,000 replications for each of the 36 simulations.

within each subclass, we regress Y on a constant, T_1^A, T_2^A, and all of the covariates. Finally, we calculate the overall average causal effect as the weighted average of the within-subclass estimates.

We compare the performance of the propensity function method with that of Gaussian linear regression, where we regress Y on T_1^A, T_2^A, and all covariates. The percent reduction in bias for 2×2, 3×3, and 4×4 subclassification schemes are given in Table 3. In all cases considered here, the biases for the causal effect of duration are more than 70% smaller with the propensity function method than with the standard linear regression adjustment. The gains offered by the propensity score methods are especially large with variable treatment effects, the case often found in practice.

4.3 Data Analysis

We now turn to the observed response variable. We use the same propensity functions as in Section 4.2 and model Y within each subclass using the same two-part model as in Section 3.3, controlling for all covariates. We also fit the smooth coefficient model by letting the effects of the two treatments be two separate unknown smooth functions of both propensity functions. As in Section 3.3, we use the generalized additive models with binomial family (logistic link) and Gaussian family (identity link) and control for all the covariates as linear predictors. Finally, we compute the weighted average of the within-subclass estimates of the coefficients.

Table 4 reports the results of the methods based on two propensity functions. All methods indicate that among smokers, the two treatment variables have no significant impact on the probability of spending some money on medical care. On the other hand, they agree that the frequency of smoking increases medical expenditure significantly, whereas the duration of smoking does not. For example, an increase from one cigarette to one pack of cigarettes per day raises annual medical expenditure by about 30%. The analysis of the bivariate treatments is more informative than the analysis in Section 3.3 in that it demonstrates that the significant effect of *packyear* is attributable mostly to the frequency of smoking rather than to its duration.

Figure 4 shows the within-subclass estimates of the causal effects under the 3×3 subclassification scheme. The within-subclass standard errors are too large to distinguish the effects among the subclasses. The added structure of the smooth coefficient model allows for more powerful comparisons, however. Figure 5 illustrates contour plots of the treatment effects as

Table 4. Estimated Average Causal Effect of Increased Smoking on Medical Expenditures via the Two Propensity Functions

	3 × 3 subclasses		4 × 4 subclasses		Smooth coefficient	
	Duration	*Frequency*	*Duration*	*Frequency*	*Duration*	*Frequency*
Logistic regression						
Coefficient	−.437	.061	−.359	.026	−.419	.087
Standard error	8.096	4.752	8.789	5.470	7.654	4.426
Gaussian regression						
Coefficient	−.010	.078	.027	.068	−.022	.088
Standard error	.036	.027	.046	.032	.034	.025

NOTE: The coefficients of two treatment variables, log(duration) and log(frequency), and their standard errors are reported.

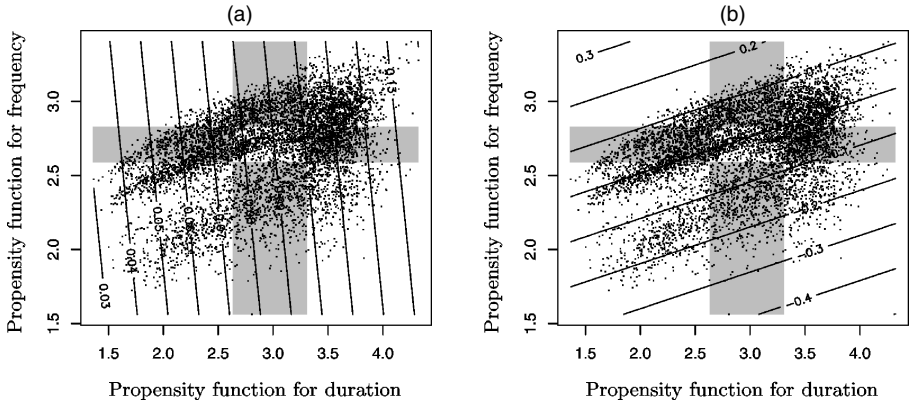

Figure 5. Estimated Causal Effects of (a) Frequency and (b) Duration From the Smooth Coefficient Model. The solid curves represent the causal effect as a function of the estimated propensity functions and are based on the estimated coefficients of log(frequency) and log(duration) from the (Gaussian) smooth coefficient model presented in Table 4. The points represent the observations, and the gray and white areas represent the nine subclasses used in the analysis based on a 3 × 3 subclassification scheme; see Figure 4.

functions of the two propensity functions; we can reject the null model of a constant treatment effect of duration with $p < .04$. Our simulation results indicates that the propensity score methods perform especially well relative to direct methods when the treatment effect varies across the population. As with the analysis of the univariate treatment in Section 3.3, it is important to remember the selection bias in the sample when interpreting the causal effects; only smokers who survive are included in the sample.

5. EFFECTS OF SCHOOLING ON INCOME

5.1 Background and Data

In this section we estimate the average causal effect of schooling on income by applying the propensity function method to balance the instruments in an instrumental variables (IV) analysis. The effect of education on income has long been an important topic in economics; researchers have quantified the effect by comparing years of education and individual wage in IV analyses (e.g., Angrist and Krueger 1991, 1992; Card 1995; Kling 2001). But the use of IV estimation in observational studies is vulnerable to criticism concerning the validity of the instrument (e.g., Bound, Jaeger, and Baker 1995). Thus improving the performance of IV estimation has been a focus of much recent literature (e.g., Angrist and Krueger 1995; Staiger and Stock 1997; Angrist, Imbens, and Krueger 1999). Here we show how the propensity function methods developed in this article can potentially be used to improve IV estimation.

Angrist and Krueger (1995) used data collected from six U.S. Current Population Surveys (CPSs) on men born between 1949 and 1953. Only the subsample of men born between 1949 to 1953 is publicly available, and we use this subsample in our analysis. Wages and other information were recorded for one of the years between 1979 and 1985 (excluding 1980); following the original article, we adjusted wages to 1978 dollars. The dataset contains nine background variables: education in terms of the highest grade completed (0–18), race (Black, Hispanic, and others), year of birth (1949–1953), marital status (single or married), veteran status (veteran or not a veteran), Vietnam lottery code (14 categories), region of residence (9 regions), and indicator variables for residence in a central city and employment in a standard metropolitan statistical area. Following Angrist and Krueger (1995), we exclude those men who did not work and/or recorded zero earnings as well as those who have missing values for at least one variable. This yields a sample size of 13,900 for our analysis.

5.2 Assumptions and Previous Analyses

Before we describe the IV analysis, we pause to consider an analysis based directly on the propensity function, that is, an analysis of the sort illustrated in Sections 3 and 4. In this case we are interested in the effect of the treatment variable, highest grade completed, on wages. The validity of the direct propensity function analysis is predicated on Assumption 2, that the treatment and the potential outcomes are independent given the set of observed covariates. Unfortunately, the set of covariates contains no measure of such important factors as underlying individual intelligence or work ethic, both of which would seem to affect the treatment and the potential outcomes. For example, individuals who are intellectually gifted and motivated tend to attain higher levels of education and might be expected to earn higher wages for any given level of education they might have attained. Without controlling for a richer set of covariates (e.g., Rouse 1995), Assumption 2 is unjustifiable. Our criticism of the ignorability assumption is substantive in nature; Rosenbaum and Rubin (1983a) described a method for quantifying the sensitivity of results to Assumption 2.

Although an IV analysis requires certain other assumptions, it does not require that the treatment assignment be ignorable. Hence an IV analysis may be more appropriate here. To estimate the causal effect of education on income, Angrist and Krueger (1995) used two-stage least squares (TSLS), a type of IV estimation. Specifically, they assumed that

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\alpha}_0 + T_i \xi + V_i \gamma + \epsilon_i, \tag{5}$$

$$T_i = \mathbf{X}_i^\top \boldsymbol{\alpha}_1 + \mathbf{Z}_i^\top \boldsymbol{\delta}_1 + u_i, \tag{6}$$

and

$$V_i = \mathbf{X}_i^\top \boldsymbol{\alpha}_2 + \mathbf{Z}_i^\top \boldsymbol{\delta}_2 + \eta_i, \tag{7}$$

where $i = 1, \ldots, n$ indexes individuals, Y_i is log weekly wage, $\mathbf{X}_i$ is a vector of covariates, T_i is the highest grade completed,

V_i is an indicator variable for veteran status, $\mathbf{Z}_i$ is a vector of IVs that interact the assigned Vietnam draft lottery code, $\tilde{Z}_i^A$, with year-of-birth indicator variables, and ϵ_i, u_i, and η_i represent independent error terms. Here ξ represents the causal effect of education on wages. The estimation procedure consists of two steps. First, the fitted values, $\widehat{T}_i$ and $\widehat{V}_i$, are obtained via the least squares fit of (6) and (7). Then T_i and V_i in (5) are replaced with their fitted values from the first step and the least squares estimate of the average treatment effect, $\hat{\xi}$, is computed.

In this formulation, the Vietnam draft lottery code plays a key role in constructing the IVs, whereas veteran status and education level form a bivariate treatment. To assign a causal interpretation to ξ, the IVs must (a) be independent of both the potential outcomes and potential treatment assignments given $\mathbf{X}$, (b) be monotonically predictive of the treatment assignment given $\mathbf{X}$, and (c) affect only the outcome variable through the treatment variables (Angrist and Imbens 1995; Angrist, Imbens, and Rubin 1996). As Angrist and Krueger (1995) pointed out, the key here is that the assignment mechanism only for the lottery code (and not that for education level) needs to be strongly ignorable. They also argued, in reference to requirement (b), that men with low draft lottery numbers, who were likely to be drafted, had a strong incentive to stay in school. Thus the key insight of the approach of Angrist and Krueger (1995) is the use of the lottery code as an instrument. [Veteran status is included in the treatment to help ensure that requirement (c) is met.]

Because the lottery code was randomly assigned, we can view this scenario as a "natural experiment." Men were randomly assigned to lottery codes; some of these codes encouraged men to go school to avoid the draft. Thus, in a sense, there are two "treatment" variables: the randomly assigned lottery code and the level of education. In this encouragement design, IV methodology allows us to estimate the causal effect of the level of education. But inference would be biased if lottery codes were correlated with the potential outcomes (i.e, income) or the potential levels of the education variable. Thus, we carry out an IV analysis using the propensity function to balance the covariates across the randomized "treatment" variable of the natural experiment, that is, the instrument.

5.3 Balancing the Covariates Across the Instrument

Although the lottery code is randomly assigned and thus the true propensity function, $p_{\boldsymbol{\psi}}(\tilde{Z}^A|\mathbf{X})$, is known and constant as a function of $\mathbf{X}$, as described in Section 2.4, adjusting for the *estimated* propensity function can still be advantageous. Briefly, a randomized treatment assignment balances the covariates only in expectation, but by adjusting for the estimated propensity function, we can bring the covariates closer to exact balance in the observed sample. This is illustrated in Figure 6 for this example. First, we regress each of the covariates on the lottery code using logistic regression. (All covariates are indicator variables.) The 22 resulting t-statistics (df = 13,898) appear in a standard normal quantile plot in Figure 6(a). There is no evidence that the lottery code is correlated with any of the covariates. The t-statistics are not 0, because the balance is not exact.

Our goal is to improve the observed balance of the IV, $\mathbf{Z}^A$, by first balancing the assigned lottery code, $\tilde{Z}^A$. (Recall that $\mathbf{Z}^A$ represents the interaction terms of $\tilde{Z}^A$ with year-of-birth indicator variable.) To do this, we condition on the estimated propensity function, $p_{\hat{\boldsymbol{\psi}}}(\tilde{Z}^A|\mathbf{X})$. In particular, we use an ordinal logistic model to estimate the conditional probability of each lottery code given all of the available covariates (see, e.g., McCullagh and Nelder 1989). Given the estimated values of the parameters, the scalar linear predictor, $\hat{\theta} = \mathbf{X}^\top \hat{\boldsymbol{\beta}}$, completely identifies the propensity function; $\boldsymbol{\beta}$ takes the role of $\boldsymbol{\psi}$ in the general framework. Figure 6(b) is identical to 6(a) except that we control for the linear predictor, $\hat{\theta} = \mathbf{X}^\top \hat{\boldsymbol{\beta}}$, in each logistic regression. The resulting t-statistics are much closer to 0, because better balance is achieved by conditioning on $\hat{\theta}$.

Taking advantage of the improved balance, we subclassify the sample on $\hat{\theta}$ into several subclasses of roughly equal size. We then replicate the TSLS analysis of Angrist and Krueger (1995) as specified in (5)–(7) within each subclass. Finally, we obtain the estimate of the average treatment effect by computing the weighted average of the within-subclass estimates. Table 5 displays the estimated average treatment effects of education, that is, the average effect of 1 additional year of education on log weekly wage. Along with the results based on TSLS and the propensity function, the table presents the estimates based on the split-sample IV (SSIV) of Angrist and Krueger (1995).

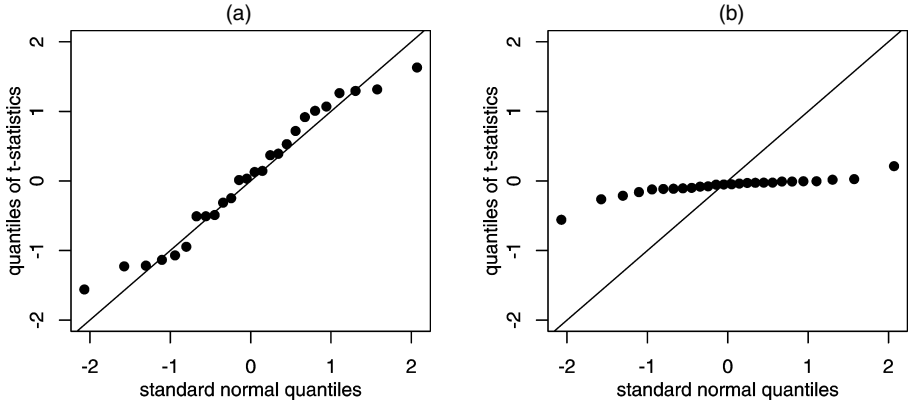

Figure 6. Standard Normal Quantile Plots of t-Statistics for the Coefficient of the Lottery Code Variable in the Models Predicting Each Covariate, for the Models That (a) Do Not Control and (b) Control for the Estimated Propensity Function.

Table 5. Estimated Average Treatment Effect of Education on Income

	Direct models		Propensity function	
	TSLS	SSIV	5 subclasses	10 subclasses
Average causal effect	.109	.040	.062	.063
Standard error	.034	.037	.015	.010

NOTE: The figures represent the average effect of a 1-year increase in the highest grade completed on log weekly wage. (See Angrist and Krueger 1995 for a complete discussion of the SSIV method.) Results for SSIV are based on 250 bootstrap samples.

Angrist and Krueger used this estimator to overcome the finite-sample bias of TSLS. They noted that SSIV estimates tend to be biased toward 0, whereas TSLS estimates tend to exhibit bias toward the least squares estimates. Balancing the instruments using the estimated propensity function reduces the TSLS estimate, but it is still not as close to 0 as the SSIV estimate.

Table 6 reports the within-subclass TSLS estimates and standard errors using five subclasses. The subclassification standard errors are smaller than those based on TSLS or SSIV.

6. CONCLUDING REMARKS

This article extends the propensity score of Rosenbaum and Rubin (1983b) along with the generalizations of Joffe and Rosenbaum (1999) and Imbens (2000) for application with general treatment regimes. In particular, our strategy allows researchers to estimate causal effects by conditioning on a low-dimensional parameterization of the the propensity function rather than on typically high-dimensional covariates. This formulation retains the powerful dimension reduction that makes propensity scores such a useful tool.

Subclassification on the propensity function can successfully reduce bias and MSE relative to standard regression techniques when analyzing the effects of general treatment regimes. Although severe model misspecification can lead to biased results, our simulation studies suggest that bias and error reduction is relatively robust to model misspecification. Because better model specifications lead to better results, however, care must be taken when selecting the model form of the propensity function and when computing the effect of the treatment conditional on the propensity function. Model diagnostics, including the examination of the resulting balance of the covariates after conditioning on the estimated propensity function, should always be thoughtfully used. As with all methods based on covariate adjustment, care must be taken to collect a sufficiently diverse class of covariates.

APPENDIX A: VERIFICATION OF RESULTS 1 AND 2

Proof of Result 1

We have

$$p\{\mathbf{T}^A|e(\cdot|\mathbf{X})\} = p(\mathbf{T}^A|\boldsymbol{\theta}) = p\{\mathbf{T}^A|\boldsymbol{\theta}(\widetilde{\mathbf{X}})\} = p(\mathbf{T}^A|\widetilde{\mathbf{X}}), \qquad \text{(A.1)}$$

for $\boldsymbol{\theta}$ such that $e(\cdot|\mathbf{X}) = e(\cdot|\boldsymbol{\theta})$, and for any $\widetilde{\mathbf{X}} \in \mathcal{X}$ such that $\boldsymbol{\theta}(\widetilde{\mathbf{X}}) = \boldsymbol{\theta}$, in particular, $\widetilde{\mathbf{X}} = \mathbf{X}$. The first equality in (A.1) follows from Assumption 3; the second, from the definition of $\boldsymbol{\theta}$; and the third, from the sufficiency of $\boldsymbol{\theta}$ for $\mathbf{T}^A$. Replacing $\widetilde{\mathbf{X}}$ with $\mathbf{X}$, this implies that the propensity function is a balancing score, because $p(\mathbf{T}^A|\mathbf{X}) = p\{\mathbf{T}^A|\mathbf{X}, e(\cdot|\mathbf{X})\} = p\{\mathbf{T}^A|e(\cdot|\mathbf{X})\}$, where the first equality follows from the fact that $e(\cdot|\mathbf{X})$ is redundant given $\mathbf{X}$.

Table 6. Within-Subclass TSLS Estimates of Average Treatment Effect of Education on Income for Each of Five Subclasses

Subclass I	Subclass II	Subclass III	Subclass IV	Subclass V
.084 (.028)	.063 (.035)	.020 (.028)	.054 (.036)	.090 (.036)

NOTE: Standard errors are given in parentheses.

Proof of Result 2

Given $e(\cdot|\mathbf{X})$, the joint distribution of $\mathbf{T}^A$, $\mathbf{X}$, and $Y(\mathbf{t}^P)$ is

$$p\{T^A, \mathbf{X}, Y(\mathbf{t}^P)|e(\cdot|\mathbf{X})\} = p\{\mathbf{T}^A, \mathbf{X}|e(\cdot|\mathbf{X})\}p\{Y(\mathbf{t}^P)|\mathbf{T}^A, \mathbf{X}, e(\cdot|\mathbf{X})\}. \qquad \text{(A.2)}$$

Applying Result 1 to factor the first term of the right-hand side of (A.2) and Assumption 2 to rewrite the second term, we have $p\{\mathbf{T}^A, \mathbf{X}, Y(\mathbf{t}^P)|e(\cdot|\mathbf{X})\} = p\{\mathbf{T}^A|e(\cdot|\mathbf{X})\}p\{\mathbf{X}|e(\cdot|\mathbf{X})\}$ $p\{Y(\mathbf{t}^P)|\mathbf{X}, e(\cdot|\mathbf{X})\}$. Combining the final two terms of this expression and integrating over $\mathbf{X}$, we find that given $e(\cdot|\mathbf{X})$, $Y(\mathbf{t}^P)$, and $\mathbf{T}^A$ are independent.

APPENDIX B: DIAGNOSTICS OF A LINEAR REGRESSION PROPENSITY FUNCTION

If a linear regression is used to model the dependence of the treatment variable on a set of covariates, then the treatment variable is necessarily uncorrelated with each covariate given the linear predictor. Although this is an indication that each covariate is balanced, the partial correlations are not useful as diagnostics of the model specification for the propensity function. This is formalized in the following result.

Result 3. Consider a full rank set of covariates, $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1, \ldots, \mathbf{X}_p)$ and a treatment variable, $\mathbf{T}$, where $\mathbf{T}$ is an $n \times 1$ vector, $\mathbf{1}$ is an $n \times 1$ vector of 1's, and $\mathbf{X}_p$ is an $n \times 1$ vector covariate for each p. Let $\widehat{\mathbf{T}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{T}$ be the linear predictor of $\mathbf{T}$. The partial correlation of $\mathbf{T}$ with each $\mathbf{X}_p$ is 0 given $\widehat{\mathbf{T}}$, that is, the second component of $(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\mathbf{T}$ is 0, where $\widetilde{\mathbf{X}} = (\mathbf{1}, \mathbf{X}_p, \widehat{\mathbf{T}})$.

Proof. If we substitute $\widehat{\mathbf{T}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{T}$ into $(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\mathbf{T}$ and use the identities $\mathbf{1}^\top\widehat{\mathbf{T}} = \mathbf{1}^\top\mathbf{T}$, $\mathbf{X}_p^\top\widehat{\mathbf{T}} = \mathbf{X}_p\mathbf{T}$, and $\widehat{\mathbf{T}}^\top\widehat{\mathbf{T}} = \widehat{\mathbf{T}}^\top\mathbf{T}$, then the result follows from algebraic manipulations.

[Received November 2002. Revised January 2004.]

REFERENCES

Angrist, J. D., and Imbens, G. W. (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999), "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14, 57–67.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association*, 91, 444–455.

Angrist, J. D., and Krueger, A. B. (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106, 979–1014.

——— (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables With Moments From Two Samples," *Journal of the American Statistical Association*, 87, 328–336.

——— (1995), "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics*, 13, 225–235.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995), "Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Card, D. E. (1995), "Earnings, Schooling, and Ability Revisited," *Research in Labor Economics*, 14, 23–48.

D'Agostino, R. B., Jr., and Rubin, D. B. (2000), "Estimating and Using Propensity Scores With Partially Missing Data," *Journal of the American Statistical Association*, 95, 451, 749–759.

Dehejia, R. H., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.

DiNardo, J., and Tobias, J. L. (2001), "Nonparametric Density and Regression Estimation," *Journal of Economic Perspectives*, 15, 11–28.

Drake, C. (1993), "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect," *Biometrics*, 49, 1231–1236.

Duan, N., Manning, W. G. J., Morris, C. N., and Newhouse, J. P. (1983), "A Comparison of Alternative Models for the Demand for Medical Care," *Journal of Business & Economic Statistics*, 1, 115–126.

Efron, B., and Feldman, D. (1991), "Compliance as an Explanatory Variable in Clinical Trials" (with discussion), *Journal of the American Statistical Association*, 86, 9–17.

Gerber, A. S., and Green, D. P. (2000), "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment," *American Political Science Review*, 94, 653–663.

Hastie, T. J., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman & Hall.

Heckman, J. J., Ichimura, H., and Todd, P. (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Hill, J., Rubin, D. B., and Thomas, N. (1999), "The Design of the New York School Choice Scholarship Program Evaluation," in *Research Designs: Inspired by the Work of Donald Campbell*, ed. L. Bickman, Thousand Oaks, CA: Sage, pp. 155–180.

Hirano, K., Imbens, G., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1307–1338.

Holland, P. W. (1986), "Statistics and Causal Inference" (with discussion), *Journal of the American Statistical Association*, 81, 945–960.

Imai, K. (2004), "Do Get-Out-the-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments," *American Political Science Review*, to appear.

Imai, K., and van Dyk, D. A. (2004), "A Bayesian Analysis of the Multinomial Probit Model Using Marginal Data Augmentation," *Journal of Econometrics*, to appear.

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

Imbens, G. W., and Rubin, D. B. (1997), "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Economic Studies*, 64, 555–574.

Javaras, K. N., and van Dyk, D. A. (2003), "Multiple Imputation for Incomplete Data With Semicontinuous Variables," *Journal of the American Statistical Association*, 98, 703–715.

Joffe, M. M., and Rosenbaum, P. R. (1999), "Propensity Scores," *American Journal of Epidemiology*, 150, 327–333.

Johnson, E., Dominici, F., Griswold, M., and Zeger, S. L. (2003), "Disease Cases and Their Medical Costs Attributable to Smoking: An Analysis of the National Medical Expenditure Survey," *Journal of Econometrics*, 112, 135–151.

Kling, J. R. (2001), "Interpreting Instrumental Variables Estimates of the Returns to Schooling," *Journal of Business & Economic Statistics*, 19, 358–364.

Larsen, M. D. (1999), "An Analysis of Survey Data on Smoking Using Propensity Scores," *Sankhyā*, Ser. B, 61, 91–105.

Lechner, M. (1999), "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification," *Journal of Business & Economic Statistics*, 17, 74–90.

Li, Q., Huang, C. J., Li, D., and Fu, T.-T. (2002), "Semiparametric Smooth Coefficient Models," *Journal of Business & Economic Statistics*, 20, 412–422.

Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. R. (2001), "Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse," *Journal of the American Statistical Association*, 96, 1245–1253.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.

Olsen, M. K., and Schafer, J. L. (2001), "A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data," *Journal of the American Statistical Association*, 96, 730–745.

Robins, J. M., and Rotnitzky, A. (2001), Comment on "Inference for Semiparametric Models: Some Questions and an Answer," by J. P. Bickel and J. Kwon, *Statistica Sinica*, 11, 920–936.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121.

Rosenbaum, P. R. (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387–394.

Rosenbaum, P. R., and Rubin, D. B. (1983a), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society*, Ser B, 45, 212–218.

——— (1983b), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

——— (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.

——— (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 39, 33–38.

Rouse, C. E. (1995), "Democratization or Diversion? The Effect of Community Colleges on Educational Attainment," *Journal of Business & Economic Statistics*, 13, 217–224.

Rubin, D. B. (1973), "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 159–183.

——— (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318–328.

——— (1980), Comments on "Randomization Analysis of Experimental Data: The Fisher Randomization Test," by D. Basu, *Journal of the American Statistical Association*, 75, 591–593.

——— (1990), Comments on "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," by J. Splawa-Neyman, translated from the Polish and edited by D. M. Dabrowska and T. P. Speed, *Statistical Science*, 5, 472–480.

——— (2000), "Statistical Issues in the Estimation of the Causal Effects of Smoking Due to the Conduct of the Tobacco Industry," in *Statistical Science in the Courtroom*, ed. J. L. Gastwirth, New York: Springer-Verlag, pp. 321–351.

——— (2001), "Estimating the Causal Effect of Smoking," *Statistics in Medicine*, 20, 1395–1414.

Rubin, D. B., and Thomas, N. (1992), "Affinely Invariant Matching Methods With Ellipsoidal Distributions," *The Annals of Statistics*, 20, 1079–1093.

——— (1996), "Matching Using Estimated Propensity Scores: Relating Theory to Practice," *Biometrics*, 52, 249–264.

——— (2000), "Combining Propensity Score Matching With Additional Adjustments for Prognostic Covariates," *Journal of the American Statistical Association*, 95, 573–585.

Staiger, D., and Stock, J. H. (1997), "Instrumental Variables Regression With Weak Instruments," *Econometrica*, 65, 557–586.

Wood, S. (2003), "Thin Plate Regression Splines," *Journal of the Royal Statistical Society*, Ser. B, 65, 95–114.

Yatchew, A. (1998), "Nonparametric Regression Techniques in Economics," *Journal of Economic Literature*, 36, 669–721.

Zeger, S., Wyant, T., Miller, L., and Samet, J. (2000), "Statistical Testimony on Damages in Minnesota versus the Tobacco Industry" in *Statistical Science in the Courtroom*, ed. J. L. Gastwirth, New York: Springer-Verlag, pp. 303–320.