# Causal Inference with General Treatment Regimes: Generalizing the Propensity Score*

Kosuke Imai
Department of Politics, Princeton University, Princeton, NJ 08544 USA.

David A. van Dyk
Department of Statistics, Harvard University, Cambridge, MA 02138 USA.

July 8, 2003

**Abstract**

In this article, we develop the theoretical properties of the propensity function which is a generalization of the propensity score of Rosenbaum and Rubin (1983b). 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 non-random 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 were suggested by Joffe and Rosenbaum (1999) and Imbens (2000) for ordinal and categorical treatments, respectively. In this article, we develop theory and methods which 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 data sets; we estimate the effect of smoking on medical expenditure and the effect of schooling on wages. We also conduct Monte Carlo experiments to investigate the performance of our methods.

Key Words: causal inference, income, medical expenditure, non-random treatment, observational studies, schooling, smoking, subclassification.

*Kosuke Imai is Assistant Professor, Department of Politics, Princeton University (Email: kimai@Princeton.Edu, WWW: http://www.princeton.edu/~kimai); and David A. van Dyk is Associate Professor, Department of Statistics, Harvard University (Email: vandyk@stat.harvard.edu). The authors thank Joshua Angrist, Guido Imbens, Elizabeth Johnson, and Scott Zegar for providing the data sets used in this article. They also thank Samantha Cook, Dan Ho, Gary King, Donald Rubin, and Elizabeth Stuart for helpful discussions. The comments from the associate editor and anonymous referees significantly improved this article. Research support was partially provided by NSF grant DMS-01-04129, by the U.S. Census Bureau, and by the Center for Basic Research in the Social Sciences at Harvard University.

# 1 Introduction

Establishing the effect of a treatment that is not randomly assigned is a common goal in empirical research. The lack of random assignment, however, 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, the causal effect of the treatment may be difficult to ascertain. 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 recent years in the social sciences (e.g. Heckman *et al.*, 1998; Lechner, 1999; Imai, 2003).

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 pre-treatment 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, 2003).

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 (strongly ignorable treatment assignment), 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 or linear discriminant analysis (Rosenbaum and Rubin, 1984, 1985). The effects of using estimated propensity scores in place of true propensity scores are discussed at length in the literature (e.g., Rosenbaum, 1987; Robins *et al.*, 1995; Rubin and Thomas, 1996; Heckman *et al.*, 1998; Hirano *et al.*, 2002); see also Section 5.3. One of the principle advantages of this method is that adjusting for the propensity score amounts to matching or subclassifying on a scalar quantity, which is

significantly easier than matching or subclassifying on a large number of 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. In many observational studies, however, 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). It is also possible that the treatment is measured in frequency and duration, e.g., the health effects of smoking. These examples taken together illustrate the need for the propensity score, a prominent methodology of causal inference, to be extended for application with 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 while a categorical variable is discrete with possibly unordered values.) Imbens (2000) suggests computing a propensity score for each level of a categorical treatment variable, i.e., 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 levels of the treatment. For an ordinal treatment variable, Joffe and Rosenbaum (1999) proposed and Lu *et al.* (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. (At the end of their paper, however, Joffe and Rosenbaum (1999) propose the possibility of adjusting for a low-dimensional linear propensity score in the context of a univariate ordinal

treatment. Imbens (2000) also suggests adjusting for several propensity scores; but the scores are adjusted for one at a time.)

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

The rest of the article is organized into five sections. Section 2 describes the propensity function which is our generalization of the propensity score. A set of Monte Carlo experiments are provided in Section 3. Sections 4 and 5 illustrate our method through two applied examples. The final section gives concluding remarks.

# 2 Methodology and Theory

## 2.1 Framework for causal inference

Consider a simple random sample of size $n$. For $i = 1, ..., n$, we observe a $p \times 1$ vector of pre-treatment covariates, $X_i$, and the possibly multivariate value of the treatment received, $T_i^A$, as well as the value of the outcome variable associated with this treatment, $Y_i$. We adopt the common framework for causal inference, frequently referred to as the Rubin causal model (Holland, 1986). In this framework, we define a set of potential outcomes, $\mathcal{Y} = \{Y_i(t^P), t^P \in \mathcal{T} \text{ for } i = 1, \ldots, n\}$, where $\mathcal{T}$ is a set of potential treatment values and $Y_i(t^P)$ is a random variable that maps a particular potential treatment, $t^P$, to a potential outcome. We treat $t^P$ as an ordinary variable while $T_i^A$ is a random variable.

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

Assumption 1 : Stable Unit Treatment Value Assumption (Rubin, 1980, 1990). *The distribution of potential outcomes for a unit is assumed to be independent of potential*

*treatment status of another unit given the observed covariates. Formally,* $p\{Y_i(t_i^P) \,|\, T_j^A = t_j^P, X_i\} = p\{Y_i(t_i^P) \,|\, X_i\}$ *for all* $i \neq j$ *and any* $t_i^P, t_j^P \in \mathcal{T}$.

ASSUMPTION 1 excludes the possibility of interference between units and allows us to conveniently consider the potential outcomes of one unit to be conditionally independent of another unit's treatment status given the observed covariates. (Thus, we may suppress the observational index, $i$, and do so for the remainder of the article.) Because the treatment assignment mechanism in most observational studies is unknown, the conditional distribution of $T^A$ given $X$ needs to be modeled, usually parametrically. The following critical assumption allows us to model $T^A$ without conditioning on potential outcomes.

ASSUMPTION 2 : STRONG IGNORABILITY OF TREATMENT ASSIGNMENT (RUBIN, 1978). *The distribution of the actual treatment does not depend on potential outcomes given the observed covariates. Formally,* $p\{T^A \,|\, Y(t^P), X\} = p(T^A \,|\, X)$ *for all* $t^P \in \mathcal{T}$ *and* $0 < p(T^A \in \mathcal{A} \,|\, X)$ *for all* $X \in \mathcal{X}$ *and measurable sets* $\mathcal{A} \subset \mathcal{T}$.

In practice, ignorability is a non-trivial 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 $p\{Y(t^P) \,|\, X\}$ as a function of $t^P$ and for fixed $X$, or its average over the population, $p\{Y(t^P)\} = \int p\{Y(t^P) \,|\, X\}\, p(X)\, dX$, is of primary interest. The fundamental difficulty of causal inference in observational studies is that we only observe one of the potential outcomes, $Y(t^P = T^A) \in \mathcal{Y}$. Therefore, in practice, we must condition on the observed treatment assignment. Because $T^A$ and $X$ are not generally independent, however, basing inference on $p\{Y(t^P) \,|\, T^A\} = \int p\{Y(t^P) \,|\, X, T^A\}\, p(X \,|\, T^A)\, dX$ often leads to bias. The solution lies in conditioning on the observed covariates; by ASSUMPTION 2, $p\{Y(t^P) \,|\, T^A, X\} \propto p\{Y(t^P), T^A \,|\, X\} = p\{Y(t^P) \,|\, X\} p\{T^A \,|\, X\} \propto p\{Y(t^P) \,|\, X\}$. Thus, the average causal effect $E\{Y(t_1^P) - Y(t_2^P) \,|\, X\} = E\{Y(t_1^P) \,|\, T^A = t_1^P, X\} - E\{Y(t_2^P) \,|\, T^A = t_2^P, X\}$, where $t_1^P \neq t_2^P$, and we obtain valid inference conditional on $X$ even when we condition on the observed treatment assignment.

In principle, we can model $p\{Y(t^P) \mid T^A = t^P, X\}$ directly, but experience shows that even with binary treatments standard model assumptions, e.g., linearity, do not suffice and that this misspecification can strongly bias causal inference (Drake, 1993; Dehejia and Wahba, 1999). A variety of non-parametric techniques exist; matching and subclassification are commonly used. However, as the dimensionality of $X$ increases, matching and subclassification become impossible in practice. The propensity score aids statistical analysis in this regard by reducing the dimensionality of the variable that is conditioned upon to a scalar variable. In the following section, we generalize the propensity score such that it is not only applicable to arbitrary treatment regimes including continuous treatments, but also reduces the dimensionality of $X$ enough 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, i.e., $p_\psi(T^A \mid X)$, where $\psi$ parameterizes this distribution. When $T^A$ is binary the propensity function is determined by the propensity score, $p_\psi(T^A \mid X)|_{T^A=1}$, where $T^A$ is an indicator variable for the treatment.

In practice, the propensity function is unknown and the conditional distribution, $p_\psi(T^A \mid X)$ must be modeled, and the unknown parameters, $\psi$, must be estimated using, e.g., maximum likelihood. Since we treat $\psi$ as an unknown but fixed quantity $\psi$ is implicitly condition upon throughout; for clarity, we occasionally denote the dependency of distributions on $\psi$ through a subscript. This parametric model defines the propensity function, $e_\psi(\cdot \mid X) = p_\psi(\cdot \mid X)$. Misspecification of the model for the propensity function is possible, and generally leads to biased 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.

In order 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 $X \in \mathcal{X}$, there exists a finite dimensional parameter, $\theta \in \Theta$, such that $e_\psi(\cdot \mid X) = e\{\cdot \mid \theta_\psi(X)\}$ and $\int_{\mathcal{A}} e_\psi(t \mid \theta)dt = \int_{\mathcal{A}} e_\psi(t \mid \theta')dt$ for all measurable sets $\mathcal{A} \subset \mathcal{T}$ imply $\theta = \theta'$.*

*That is, $\theta$ uniquely represents $e\{\cdot \mid \theta_\psi(X)\}$, which we may therefore write as $e(\cdot \mid \theta)$.*

This assumption implies that $e(\cdot \mid X)$ depends on $X$ only through $\theta_\psi(X)$, i.e., $\theta$ is sufficient for $T^A$. In this case, the propensity function is effectively summarized by the parameter $\theta$, which is typically of much lower dimension than is $X$. To illustrate ASSUMPTION 3 and methods based on the propensity function, we consider three simple examples.

**Example with a continuous treatment:** Suppose we model the conditional distribution of the treatment given a $(p \times 1)$ vector of covariates, $X$, as $T^A \mid X \sim N(X^\top \beta, \sigma^2)$ where $\sigma^2$ is a scalar, and $\beta$ is a $(p \times 1)$ vector of regression coefficients. Thus, the propensity function, $e\{\cdot \mid \theta_\psi(X)\}$, is the Gaussian density function, $\psi = (\beta, \sigma^2)$, and $\theta_\psi(X) = X^\top \beta$. Given $\psi$, the propensity function is completely determined by the scalar, $\theta$. Hence, matching or subclassifying on the propensity function can be easily accomplished by matching or subclassifying on $\theta$, regardless of the dimension of $X$.

**Example with a categorical treatment:** In this example, we illustrate that the propensity function encompasses the propensity scores suggested by Imbens (2000) for a categorical treatment. Suppose $\mathcal{T} = \{1, \ldots, t_{\max}\}$ and we model $p_\psi(T^A|X)$ as a multinomial distribution with probability vector $\pi(X) = \{\pi_1(X), \ldots, \pi_{\max}(X)\}$. If for each $X$, $\pi(X)$ is an unconstrained probability vector, then $\theta_\psi(X) = \pi(X)$ is a $t_{\max}$ dimensional parameter which corresponds to the set of $t_{\max}$ propensity scores proposed by Imbens (2000). We might use nested logistic regression (as suggested by Imbens, 2000) or a multinomial probit model (e.g., Imai and van Dyk, 2003) to model the dependence of $\pi(X)$ on $X$; in either case $\psi$ represents the set of 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 set up as in the example with a categorical treatment, except we model $\pi(X)$ using an ordinal logistic model (McCullagh and Nelder, 1989). In this case, $\pi(X)$ is determined by the scalar $X^\top \beta$, where $\beta$ is a $(p \times 1)$ parameter vector; in the general framework $\psi = \beta$ and $\theta_\psi(X) = X^\top \beta$. Lu *et al.* (2001) mention the possibility of using Gaussian linear

regression to model the assignment mechanism for an ordinal treatment, but must assume constant residual variance. This constraint is not necessary under our general framework, but allowing for non-constant variance generally increases the dimension of $\theta_\psi(X)$.

### 2.3 Large Sample Theory

Under the analytical framework and assumptions given in Sections 2.1 and 2.2, we derive theoretical results which closely follow and extend those in Rosenbaum and Rubin (1983b). Throughout we assume the propensity function including the parameters, $\psi$, is known. First, THEOREM 1 states that the propensity function is a balancing score even with a non-binary treatment. That is, we show that given the propensity function, the conditional distribution of the actual treatment does not depend on observed covariates.

THEOREM 1 : PROPENSITY FUNCTION AS A BALANCING SCORE.

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

PROOF : We have

$$p\{T^A \,|\, e(\cdot \,|\, X)\} = p(T^A \,|\, \theta) = p\{T^A \,|\, \theta(\widetilde{X})\} = p(T^A \,|\, \widetilde{X}), \tag{1}$$

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

In practice, THEOREM 1 can be checked, for example, by examining the $t$-statistics for the coefficient of $T^A$ in linear models that predict each covariate while controlling for the estimated propensity function. In Sections 4 and 5, we employ this diagnostic of the model specification of the propensity function; see also Appendix A.

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

THEOREM 2 : STRONG IGNORABILITY OF TREATMENT ASSIGNMENT GIVEN THE PROPENSITY FUNCTION. $p\{Y(t^P) \,|\, T^A, e(\cdot \,|\, X)\} = p\{Y(t^P) \,|\, e(\cdot \,|\, X)\}$ for any $t^P \in \mathcal{T}$.

PROOF : Given $e(\cdot \,|\, X)$, the joint distribution of $T^A$, $X$, and $Y(t^P)$ can be written

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

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

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

$$p\{Y(t^P)\} \;=\; \int p\{Y(t^P) \,|\, T^A = t^P, \theta\}\; p(\theta)\; d\,\theta, \tag{3}$$

where $\theta = \theta_\psi(X)$ uniquely indexes the propensity function.

### 2.4 From Theory to Practice

We generally accomplish the integration in (3) by subclassifying similar values of $\theta$. In particular, we first model $p_\psi(T^A \,|\, X)$ and compute the estimate $\hat{\psi}$ of $\psi$, perhaps by maximum likelihood. We then compute $\hat{\theta} = \theta_{\hat{\psi}}(X)$ for each observation and subclassify observations with the same or similar values of $\hat{\theta}$ into a moderate number of subclasses of roughly equal size. Within each subclass we model $p\{Y(t^P) \,|\, T^A = t^P\}$ and compute the relevant causal effect, e.g., the regression coefficient of $Y(t^P)$ on $t^P$. In practice, additional adjustment for the estimated propensity function within each subclass is desirable to further reduce bias. That is, we model $p\{Y(t^P) \,|\, T^A = t^P, \hat{\theta}\}$, for example, by regressing $Y(t^P)$ on both $T^A$ and $\hat{\theta}$ or some transformation thereof. To further reduce bias, some authors have suggested the inclusion of covariates in this regression (e.g. Robins and Rotnitzky, 2001). Although this is a useful strategy in some cases, we suppress such conditioning in our general notation.

The average causal effect can be computed as a weighted average of the within-subclass effects with weight equal to the relative size of the subclasses. Formally, we approximate (3) with

$$p\{Y(t^P)\} \;=\; \int p\{Y(t^P)\,|\,T^A = t^P, \theta)\; p(\theta)\; d\,\theta \;\;\approx\;\; \sum_{j=1}^{J} p\{Y(t^P)\,|\,T^A = t^P, \hat{\theta}_j\}\, W_j, \quad (4)$$

where $J$ is the number of subclasses and $W_j$ is the relative size of the subclasses. If $W_j$ is known and the estimate of the causal effect is unbiased within each subclass, 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 $j$. Since results may be sensitive to the number of subclasses and the choice of subclassification on $\hat{\theta}$, we suggest conducting a sensitivity analysis, repeating the analysis with different subclassification schemes.

Equation (4) describes how we can approximate the full distribution of the potential outcome at a particular level of the treatment. Although this full distribution is sometimes appropriate in practice (e.g., Imbens and Rubin, 1997), more often it is summarized by its mean. This is the approach we take in our examples, i.e., we compute

$$E\{Y(t^P)\} \;\approx\; \sum_{j=1}^{J} E\{Y(t^p)\,|\,T^A = t^P, \hat{\theta}_j\} W_j \quad (5)$$

In contrast to our general strategy of subclassifying on $\hat{\theta}$, Lu *et al.* (2001) suggest matching pairs of units on $\hat{\theta}$. As Lu *et al.* (2001) point out, however, matching is inherently more difficult when the treatment is not binary. In particular, not only should the matched pairs have similar values of $\hat{\theta}$, but they should also have dissimilar treatments; this second concern does not arise with a binary treatment since each pair consists of a unit from the treatment group and a unit from the control group. To accomplish matching, Lu *et al.* (2001) propose a distance measure that decreases both as the propensity scores become similar and as the assigned treatments become dissimilar.

With matched pairs in hand, Lu *et al.* (2001) suggest evaluating the treatment effect by examining the difference in response between the "high" and "low" treatments; the treatment was an ordinal variable. Because this approach ignores the magnitude of the difference in treatment, they also suggest regressing the difference in response on the difference in treatment. Although these suggestions are quite reasonable in their application, they are difficult

to generalize to unordered treatments or combinations of treatments. In practice, we may want to allow for more complex non-linear relationship between treatment and response. For example, the response variable in the study of Lu *et al.* (2001) consisted of four ordinal variables. One might wish to model the response with a ordinal logit model, perhaps accounting for correlation among the four variables. Although such models are straightforward to fit within subclasses, they require sophisticated analysis to fit with matched pairs. Of course, this situation is even more complex when $\theta_\psi(X)$ is not a scalar. Thus, although matching methods may be useful in particular settings and certainly deserve further study for general treatment regimes, we believe subclassification is a more generally applicable strategy because it allows for simpler implementation of more complex analysis models.

# 3 Monte Carlo Experiments

In this section, we use two Monte Carlo experiments to illustrate how controlling for the propensity function can improve the statistical properties of estimated causal effects.

## 3.1 A Univariate Continuous Treatment

We begin with an experiment involving the model for a continuous treatment variable that we introduced at the end of Section 2.2. In particular, we generated 5,000 data sets, each of size 1,000. For each unit within each data set, we independently draw two covariates, $X_1$ and $X_2$, from independent univariate Gaussian distributions with unit variance and means equal to one and two respectively. Then, we simulate the treatment variable, $T^A$, which depends on $X_1$ and $X_2$ through a univariate Gaussian distribution, namely

$$T^A \,|\, X_1, X_2 \overset{\text{indep.}}{\sim} N\left(1 + X_1 X_2 + X_1^2 + X_2^2,\ 1\right). \tag{6}$$

Similarly, the outcome variable, $Y$, is generated from another univariate Gaussian distribution given $X_1, X_2$ and $T^A$,

$$Y \,|\, T^A, X_1, X_2 \overset{\text{indep.}}{\sim} N\left(1 + T^A + X_1 X_2 + X_1^2 + X_2^2,\ 1\right). \tag{7}$$

In this experiment, the conditional mean of $Y$ is not linear in $X_1$ or $X_2$, and, on average given $X_1$ and $X_2$ the causal effect of a one unit change in $T^A$ is a one unit change in $Y$.

| | Average Causal Effect of $T^A$ | |
|---|---|---|
| | Bias | MSE |
| Linear Regression | 0.832 | 0.692 |
| Propensity Function | 0.390 | 0.153 |

Table 1: Performance of subclassification on the estimated propensity function compared with linear regression. The two columns represent the bias and mean squared error (MSE) based on 5,000 simulations.

For each data set, we estimate the model, $p(T^A \mid X)$, with a Gaussian linear regression with mean $X^\top\beta$ and constant variance, $\sigma^2$, where $X = (1, X_1, X_2)^\top$, $\sigma^2$ is a scalar, and $\beta$ is a $3 \times 1$ vector of fixed coefficients. We intentionally use a misspecified model in order to investigate the effect of misspecifying the propensity function on the resulting estimates. We obtain the maximum likelihood estimates for the parameter, $\theta = X^\top\beta$, that uniquely defines this propensity function. Using the estimated propensity function, we subclassify the observations into ten subclasses of roughly equal size. We then regress $Y$ on the treatment variable, $T^A$, the propensity function, via $\hat{\theta} = X^\top\hat{\beta}$, and an intercept within each subclass. This allows us to estimate the within-subclass average causal effect. Finally, we average the ten within-subclass estimates to get the overall average causal effect.

Table 1 compares the performance of subclassification on the estimated propensity function with the direct Gaussian linear regression of $Y$ on $T^A$ and $X$. Note that this direct regression is misspecified in the same manner as the model for the propensity function so that we can examine the relative sensitivity of the two methods to model misspecification. Subclassification on the propensity function significantly improves the regression estimate; it reduces the bias by over 50 percent. The propensity function method also has much smaller mean squared error. This example illustrates that subclassification on the propensity function can successfully reduce bias and improve efficiency. It also illustrates that the propensity function method can be more robust to model misspecification than direct linear regression (see Drake (1993) for similar results with a binary treatment). As pointed out by a referee, subclassification (or matching) is required to realize this gain; if we were to regress the response on $T^A$ and $\hat{\theta}$ without subclassification, the estimated causal effect will be identical to the direct regression estimate.

### 3.2 Multiple Treatments

As a second example, we consider the common and important situation where the effects of multiple treatments are of interest. Such multiple treatments are frequently encountered in applied research. For example, medical researchers may use propensity scores to adjust for non-compliance in an randomized experiment, where the combined effect of two different drugs are of interest. We illustrate the estimation of multiple treatment effects in Section 4 where we directly analyze the effects of frequency and duration of smoking on medical expenditure.

Here, we investigate the application of the propensity function to multiple treatments using a Monte Carlo experiment. In particular, we extend the experiment in Section 3.1 by adding a second continuous treatment, $T^{A_2}$, which we simulate from a bivariate Gaussian distribution together with $T^{A_1}$ for each unit in the sample,

$$\begin{pmatrix} T^{A_1} \\ T^{A_2} \end{pmatrix} \Bigg| X_1, X_2 \overset{\text{indep.}}{\sim} N_2 \left\{ \begin{pmatrix} 1 + X_1^2 + X_2^2 \\ 1 + X_1 X_2 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right\}. \tag{8}$$

Since we may be interested in the interaction effect as well as main effects, the outcome variable, $Y$, is generated from a univariate Gaussian distribution, which includes an interaction

$$Y \,|\, T^{A_1}, T^{A_2}, X_1, X_2 \overset{\text{indep.}}{\sim} N\left(1 + T^{A_1} + T^{A_2} + T^{A_1} T^{A_2} + X_1 X_2 + X_1^2 + X_2^2,\ 1\right). \tag{9}$$

In this setup, the true main causal effects of the treatments, $T^{A_1}$ and $T^{A_2}$, as well as the true interaction effect of the two treatments are all equal to one. We generated 5,000 data sets according to this model, each with sample size $n = 2,000$.

We begin by estimating the propensity function for each treatment. For $T^{A_1}$, we use the same misspecified model as we used in Section 3.1. Namely, we use a Gaussian linear regression with mean $X^\top \beta$ and constant variance, $\sigma^2$, where $X = (1, X_1, X_2)^\top$ and $\sigma^2$ is a scalar. The same misspecified model is used to independently model the propensity function for $T^{A_2}$.

Given the estimated propensity functions for the two treatments summarized by $\hat{\theta}_1 = X^\top \hat{\beta}_1$ and $\hat{\theta}_2 = X^\top \hat{\beta}_2$, respectively, we subclassify all of the observations into nine separate subclasses. Each subclass contains units with a specific range of both $\hat{\theta}_1$ and $\hat{\theta}_2$. As Figure 1

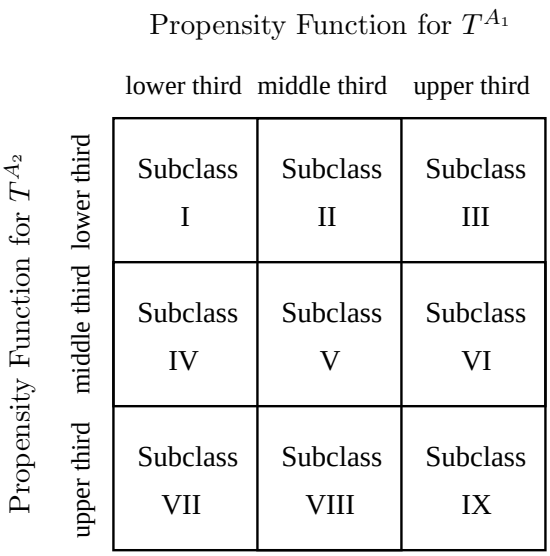


Figure 1: Subclassification of the propensity function for two overlapping treatments. Each cell of the $3 \times 3$ table represents a subclass within which units have a particular range of the propensity functions for the two treatments, $T^{A_1}$ and $T^{A_2}$. The vertical and horizontal lines which divide each subclass are the 33rd and 67th percentile of the propensity functions, as measured by $\hat{\theta}_1$ and $\hat{\theta}_2$.

illustrates, in the $3 \times 3$ 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 to maintain subclasses that are of roughly equal size.) Next, we estimate the average causal effects within each subclass using Gaussian linear regression. Namely, within each subclass, we regress $Y$ on a constant, $T^{A_1}$, $T^{A_2}$, the interaction term, $T^{A_1}T^{A_2}$, $\hat{\theta}_1$ and $\hat{\theta}_2$. Finally, the overall average causal effect is calculated as the weighted average of the nine within-subclass estimates.

Table 2 displays the simulation results. We compare the performance of the propensity function method with that of Gaussian linear regression, where we regress $Y$ on $T^{A_1}$, $T^{A_2}$, $T^{A_1}T^{A_2}$, $X_1$, and $X_2$. Notice that for the purpose of comparison this model is misspecified in a way similar to the two propensity function models. The subclassification on the propensity function results in substantially smaller bias and mean squared error than the direct regression. For example, the biases for the main causal effects of $T^{A_1}$ and $T^{A_2}$ are

| | Effect of $T^{A_1}$ | | Effect of $T^{A_2}$ | | Effect of $T^{A_1}T^{A_2}$ | |
|---|---|---|---|---|---|---|
| | Bias | MSE | Bias | MSE | Bias | MSE |
| Direct Regression | 0.517 | 0.268 | −0.340 | 0.118 | 0.045 | 0.002 |
| Propensity Function | 0.159 | 0.026 | −0.102 | 0.014 | −0.007 | 0.000 |

Table 2: Performance of subclassification on the estimated propensity function compared with direct regression. The estimates for the propensity function method are the weighted average of the nine within-subclass average causal effects. The results for the direct model are obtained from Gaussian linear regression applied to the entire sample. The columns report bias and mean squared error (MSE), and are based on 5,000 simulations.

about 70 percent smaller with the propensity function method than with the standard linear regression adjustment. Moreover, the mean square error of the propensity function method is much smaller than that of the direct regression.

The simulation studies in this section are designed to show that subclassifying on the estimated propensity function is more robust to possible model misspecification than is linear regression. Although our simulations are clearly limited in scope, we expect that as long as it effectively balances important covariates (a property that we can easily check) adjusting for the propensity function will generally lead to robust methods. This is because the within-subclass models are relatively simple in that they need not specify the relationship between the typically high-dimensional covariates and the response.

# 4 Effects of Smoking on Medical Expenditures

## 4.1 Background, Data, and Previous Studies

As a first applied example, we estimate the average effect of smoking and the amount of smoking on annual medical expenditures for individuals. Being 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 *et al.*, 2000, and the references therein). The lack of experimental data led many researchers to use the propensity score method. However, since the propensity score method is confined to a binary treatment, the focus has been on the comparison of smokers and non-smokers without distinguishing among

smokers based on how much they smoke (e.g., Larsen, 1999; Rubin, 2001). In contrast, the method we propose can incorporate the frequency and duration of smoking, non-binary treatment variables, and can be used to estimate their causal effects on health and medical expenditure.

We use the data collected by Johnson *et al.* (2003) who extracted relevant information from the 1987 National Medical Expenditure Survey (NMES, US Department of Health and Human Services). The advantages of the NMES are that it includes detailed information about frequency and duration of smoking, and that medical costs for 1987 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 in Johnson *et al.* (2003), we conduct a complete-case analysis by discarding all individuals with nonresponse. Johnson *et al.* (2003) note that better accounting for the missing data using multiple imputation did not significantly affect their results. In general, the 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, since our purpose is to illustrate the use of the propensity function, we focus on the complete-case analysis.

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. Specifically, the authors first modeled the "smoking-attributable fraction of disease" for diseases whose predominant cause is known to be smoking. These diseases include lung and laryngeal cancer and coronary heart disease. Next, they modeled the disease-attributable fraction of medical expenditures using the probability of having one of the diseases as the propensity score. In contrast, we will directly estimate the effects of smoking on medical expenditures.

### 4.2 A Continuous Scalar Treatment Variable

Previous studies have mainly focused on comparing smokers with non-smokers to estimate the direct effects of smoking (e.g., Larsen, 1999; Rubin, 2001). We focus on smokers and begin with a measure of the cumulative exposure to smoking in order to differentiate among smokers according to how much they have smoked. 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 is called *packyear* and is defined as

$$packyear \;=\; \frac{\text{number of cigarettes per day}}{20} \;\times\; \text{number of years smoked.} \qquad (10)$$

In our initial analysis, we use $\log(packyear)$ as the treatment variable.

To apply the propensity function method, we use Gaussian linear regression to model the treatment variable, $T^A = \log(packyear)$, given all available covariates, $X$; we use the sampling weights provided with the data set when fitting this regression model. The complete-case sample size for our analysis is 9,073 smokers. The estimated propensity function in this case is uniquely defined by the fitted values of $T^A$ under the model, i.e., $\hat{\theta} = X^\top \hat{\beta}$, where $\hat{\beta}$ is the maximum likelihood estimates of regression coefficients.

To evaluate the balance of the covariates, we regress each covariate on the treatment variable, $\log(packyear)$, using (unweighted) logistic linear and Gaussian linear regression for indicator and continuous covariates, respectively. (We use the log transformation of continuous covariates because $\log(packyear)$ and each covariate are necessarily uncorrelated given $\hat{\theta}$; see Appendix A.) The left panel of Figure 2 presents a standard normal quantile plot of the $t$-statistics (d.f.$= 9,071$) 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. The right panel of Figure 2 is identical to the left panel except that we control for the estimated propensity function as quantified by $\hat{\theta}$ in each regression. The figure shows the substantial reduction in the $t$-statistics that is obtained by conditioning on the estimated propensity function, indicating that the covariate balance is significantly improved. The quantile plots in Figure 2 are constructed including the square terms of the age variables (the current age and the age when the individual started smoking). The inclusion of these variables improves the balance of the covariates. In particular, if these

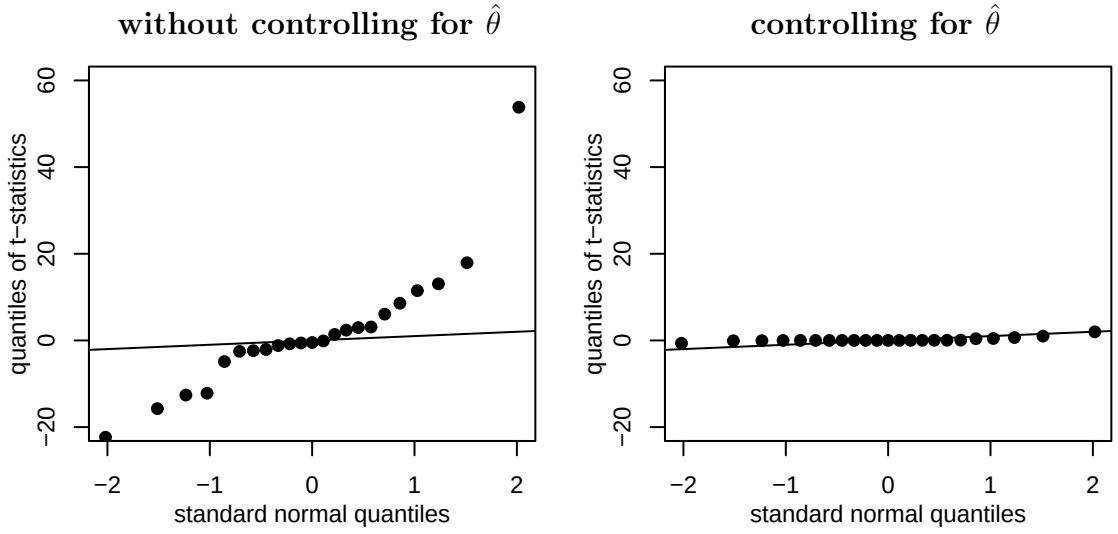


Figure 2: Standard normal quantile plots of $t$-statistics for the coefficient of $\log(\textit{packyear})$ in the models predicting each covariate. The left and right panels represent the models with and without controlling for the linear predictor of the estimated propensity function.

variables are not included, the $t$-statistic for $\log(\textit{packyear})$ as a 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 0.40. This is an example of a check of the model specification for the propensity function that was suggested in Sections 2.2 and 2.3.

Using the estimated propensity score based on this model, we subclassify the observations into ten subclasses of roughly equal size. The outcome variable is self-reported medical expenditure, denoted by $Y$, and is modeled within each subclass using the two-part model of Duan *et al.* (1983) for semi-continuous variables; see Johnson *et al.* (2003). In particular, we first model the probability of spending some money on medical care, $\Pr(Y > 0 \,|\, T^A, \hat{\theta})$, given the treatment variable, $T^A = \log(\text{packyear})$, and the linear predictor, $\hat{\theta}$ using logistic regression. Second, we model the conditional distribution of $\log(Y)$ given $T^A$ and $\hat{\theta}$, $p(\log(Y) \,|\, Y > 0, T^A, \hat{\theta})$, for those individuals who reported positive medical expenditure using Gaussian linear regression; see Olsen and Schafer (2001) and Javaras and van Dyk (2003) for discussion of models for semi-continuous variables. Using this two-part model, we estimate the effects of smoking on medical costs within each of the ten subclasses. Finally,

| | | Propensity Function | |
|---|---|---|---|
| | Direct Models | 3 subclasses | 10 subclasses |
| Logistic Linear Regression Model | | | |
| coefficient for $T^A$ | $-0.097$ | $-0.069$ | $-0.073$ |
| standard error | 3.074 | 3.035 | 3.072 |
| Gaussian Linear Regression Model | | | |
| average causal effect | 0.026 | 0.049 | 0.051 |
| standard error | 0.016 | 0.017 | 0.018 |

Table 3: Estimated average causal effect of increased smoking on medical expenditures. The logistic linear regression model presents the coefficient of the treatment variable from the models for $\Pr(Y > 0 \,|\, T^A, X)$ and $\Pr(Y > 0 \,|\, T^A, \hat{\theta})$. The Gaussian linear regression model presents the estimated average causal effects of $\log(\textit{packyear})$ on $\log(\textit{medical expenditure})$ for those individuals who reported positive medical costs (calculated from the models for $p\{\log(Y) \,|\, Y > 0, T^A, X\}$ and $p\{\log(Y) \,|\, Y > 0, T^A, \hat{\theta}\}$). For the propensity function method, the results from subclassification with three and ten subclasses are presented.

we compute the weighted average of the ten within-subclass estimates to obtain the average causal effect; in each within-subclass analysis, we use the sampling weights provided in the data set. The analysis was repeated with three subclasses, but this had little effect on the estimates.

Table 3 presents the results from subclassification with the propensity function as well as the results based on the standard complete-case linear and logistic regressions. All of the covariates are included in the standard regression models. Both methods indicate that cumulative exposure to smoking has no significant effect on the probability of spending some money on medical care in 1987. In contrast, we find that smoking, as measured by the *packyear* variable, 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, referred to as the *death benefit*.) Moreover, the propensity function method yields a greater effect of smoking on medical expenditure than the standard linear regression analysis. In particular, if *packyear* doubles we expect annual medical expenditure to increase by a factor of about 1.04.

| | $3 \times 3$ subclasses | | $4 \times 4$ subclasses | |
|---|---|---|---|---|
| | $T_1^A$ | $T_2^A$ | $T_1^A$ | $T_2^A$ |
| Logistic Linear Regression Model | | | | |
| coefficient | $-0.358$ | 0.075 | $-0.359$ | 0.026 |
| standard error | 7.110 | 4.527 | 8.789 | 5.470 |
| Gaussian Linear Regression Model | | | | |
| average causal effect | 0.011 | 0.084 | 0.027 | 0.068 |
| standard error | 0.036 | 0.026 | 0.046 | 0.032 |

Table 4: Estimated average causal effect of increased smoking on medical expenditures via the subclassification on two propensity functions. The coefficients of two treatment variables, $T_1^A$ =log(duration) and $T_2^A$=log(frequency), and their standard errors are reported.

### 4.3 A Continuous Bivariate Treatment Variable

Instead of combining frequency and duration into a single measure, we can conduct an analysis with a bivariate treatment with one variable for each characteristic. To do this, we must estimate two propensity functions, one for the frequency of smoking (the log number of cigarettes per day) and one for the duration of smoking (the log number of smoking years); we use two independent Gaussian linear regression models. In addition to the covariates, we again include the square terms for the two age variables in both models to improve the balance given the two linear predictors. We subclassify the observations into nine subclasses based on the estimated propensity functions as illustrated in Figure 1. Finally, we apply the same two-part model as in Section 4.2 within each subclass, controlling for the estimated propensity functions, and again compute the weighted average of the coefficients.

Table 4 reports the results based on the subclassification on two propensity functions. The analysis based on $3 \times 3$ and $4 \times 4$ subclasses both 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, we find that the frequency of smoking increases medical expenditure significantly while duration of smoking does not. An increase from one cigarette to one pack of cigarettes per day raises annual medical expenditure by about 30%.

# 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). The use of IV estimation in observational studies, however, is vulnerable to criticism concerning the validity of the instrument (e.g., Bound *et al.*, 1995). Thus, improving the performance of the IV estimation has been a focus of much recent literature (e.g., Angrist and Krueger, 1995; Staiger and Stock, 1997; Angrist *et al.*, 1999). Here, we show how the propensity function methods developed in this article can potentially be used to improve IV estimation.

We analyze a data set used in Angrist and Krueger (1995) that contains a sample of 16,193 individual men from six U.S. Current Population Surveys (CPS). The men were born between 1949 and 1953, and their 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. We use a subsample of the data used in the Angrist and Krueger study. They used the sample of men born between 1944 and 1953, but, only the 1949 to 1953 subsample is publicly available. In addition, the data set contains nine background variables: education in terms of the highest grade completed (0 – 18), race (Black, Hispanics, and others), year of birth (1949 – 53), 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, i.e., an analysis of the sort illustrated in Section 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 upon 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) describe a method to quantify the sensitivity of results to ASSUMPTION 2.

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

$$Y_i = X_i^\top \alpha_0 + T_i \xi + V_i \gamma + \epsilon_i, \tag{11}$$

$$T_i = X_i^\top \alpha_1 + Z_i^\top \delta_1 + u_i, \tag{12}$$

$$V_i = X_i^\top \alpha_2 + Z_i^\top \delta_2 + \eta_i, \tag{13}$$

where $i = 1, \ldots, n$ indexes individuals, $Y_i$ is log weekly wage, $X_i$ is a vector of covariates, $T_i$ is the highest grade completed, $V_i$ is an indicator variable for veteran status, $Z_i$ is a vector of instrumental variables that interact the Vietnam draft lottery code with year-of-birth indicator variables, and $\epsilon_i, u_i$, and $\eta_i$ represent independent error terms. Here, $\xi$ represents the causal effect of education on wages. The estimation procedure consists of two steps. First, one obtains the fitted values, $\widehat{T}_i$ and $\widehat{V}_i$, via the least squares fit of (12) and (13), respectively. In the second step, $T_i$ and $V_i$ in (11) 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 instrumental variables, whereas veteran status and education level form a bivariate treatment. In order to assign a causal interpretation to $\xi$, the instrumental variables must be (i) independent of both the potential outcomes and potential treatment assignments given $X$, (ii) monotonically predictive of the treatment assignment given $X$, and (iii) only affect the outcome variable through the treatment variables (Angrist and Imbens, 1995; Angrist *et al.*, 1996). As Angrist and Krueger (1995) pointed out, the key here is that only the assignment mechanism for the lottery code (and not for education level) needs to be strongly ignorable. They also argue, in reference to requirement (ii), that men with low draft lottery numbers, who were likely to be drafted, had a strong incentive to stay in school longer. 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 as part of the treatment to help ensure requirement (iii) is met.)

### 5.3 Balancing the Covariates Across the Instrument

Although the lottery code is randomly assigned and thus the true propensity function, $p_\psi(Z^A \mid X)$, is known and constant as a function of $X$, adjusting for the *estimated* propensity function can still be advantageous. Indeed, there is a large literature on the advantage of adjusting for the estimated rather than the true propensity score in both observational studies (e.g., Rosenbaum, 1984, 1987) and randomized experiments (e.g., Rubin and Thomas, 1992, 1996; Hill *et al.*, 1999). Briefly, 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 data. This is illustrated in Figure 3 for the lottery codes. First, we regress each of the covariates on the lottery code, using logistic regression. (All covariates in this case are indicator variables.) The 22 resulting $t$-statistics ($d.f. = 13,898$) are represented in a standard normal quantile plot in the left panel of Figure 3; there is no evidence that the lottery code are correlated with any of the covariates. That the $t$-statistics are not zero reflects the fact that exact balance is not achieved.

Our goal is to improve the observed balance of the instrumental variable, $Z^A$, by first balancing the assigned lottery code, $Z^{*A}$. (Recall that $Z^A$ represents the interaction terms

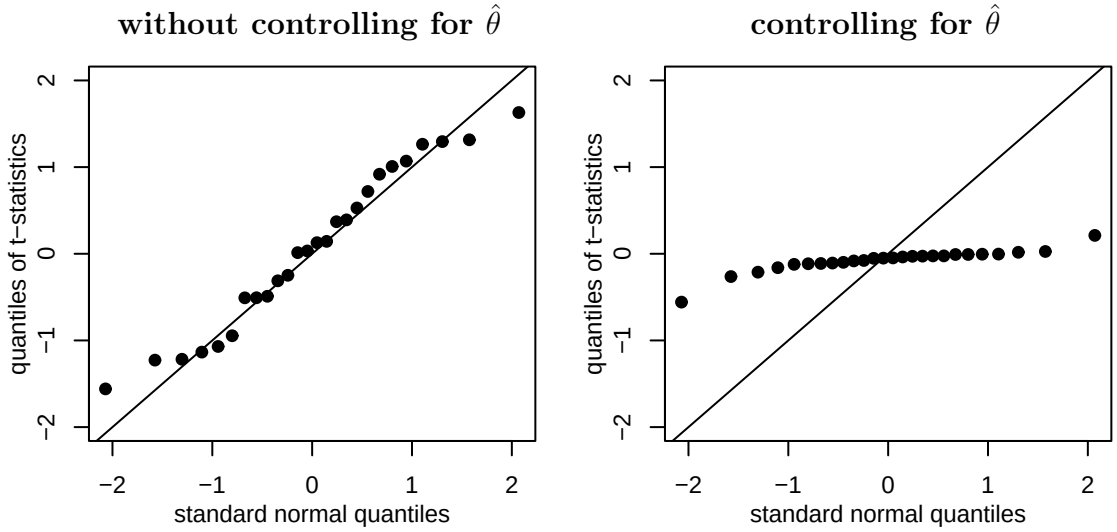


Figure 3: Standard normal quantile plots of $t$-statistics for the coefficient of the lottery code variable in the models predicting each covariate. The left panel and right represents the models with and without controlling for the linear predictor of the estimated propensity function.

of $Z^{*A}$ with year-of-birth indicator variable.) To do this, we condition on the estimated propensity function, $p_{\hat{\psi}}(Z^{*A} \mid X)$. 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, $\hat{\theta} = X^\top \hat{\beta}$, completely identifies the propensity function; $\beta$ takes the role of $\psi$ in the general framework. The second panel of Figure 3 is identical to the first except that we control for the linear predictor, $\hat{\theta} = X^\top \hat{\beta}$, in each logistic regression. The resulting $t$-statistics are all much closer to zero because better balance is achieved by conditioning on the estimated propensity function.

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 equations (11)–(13) within each subclass. When doing so, we do not further condition on $\hat{\theta}$ because $\hat{\theta} = X^\top \hat{\beta}$ is a linear function of $X$. Finally, we obtain the estimate of the average treatment effect by computing the weighted average of the within-

| | Direct Models | | Propensity Function | |
|---|---|---|---|---|
| | TSLS | SSIV | 5 subclasses | 10 subclasses |
| average causal effect | 0.109 | 0.040 | 0.062 | 0.063 |
| standard error | 0.034 | 0.037 | 0.015 | 0.010 |

Table 5: Estimated average treatment effect of education on income, i.e., the average effect of a one year increase in the highest grade completed on log weekly wage. See Angrist and Krueger (1995) for a complete discussion of the Split-Sample Instrumental Variables (SSIV) method. Results for SSIV are based on 250 bootstrap samples.

| | Subclass I | Subclass II | Subclass III | Subclass IV | Subclass V |
|---|---|---|---|---|---|
| average causal effect | 0.084 | 0.063 | 0.020 | 0.054 | 0.090 |
| standard error | 0.028 | 0.035 | 0.028 | 0.036 | 0.036 |

Table 6: Within-subclass TSLS estimates of average treatment effect of education on income for each of five subclasses. The subclassification is performed on the estimated propensity function.

subclass estimates. Table 5 displays the estimated average treatment effects of education; i.e., the average effect of one additional year of education on log weekly wage. Along with the results based on TSLS and the propensity function, we present the estimates based on the Split-Sample Instrumental Variables (SSIV) of Angrist and Krueger (1995). Angrist and Krueger used this estimator in order to overcome the finite sample bias of TSLS. They note that SSIV estimates tend to be biased toward zero whereas TSLS estimates tend to exhibit bias toward the least squares estimates. Our analysis shows that balancing the instruments using the estimated propensity function method reduces the TSLS estimate, but it is still not as close to zero as the SSIV estimate.

Table 6 reports the within-subclass TSLS estimates and standard errors using five subclasses. That the within-subclass estimates vary significantly among the subclasses illustrates the advantage of subclassification on $\hat{\theta}$. The standard errors for the estimates which balance the covariates using the estimated propensity function 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 mean squared error relative to standard regression techniques when analyzing the effects of general treatment regimes. Our simulation studies indicate that this bias and error reduction is relatively robust to model misspecification. While severe model misspecification can lead to biased results, it appears that appropriate subclassification on the propensity function reduces this bias relative to standard regression methods. Since 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 employed. As with all methods based on covariate adjustment, care must be taken to collect a sufficiently diverse class of covariates.

Nevertheless, our strategy offers advantages over other methods. For example, usual goodness-of-fit diagnostics of the standard regression analysis provide little insight as to whether a model is appropriate for causal inference. The propensity function method, on the other hand, provides guidelines about how to appropriately control for and balance a large class of covariates as is often necessary for causal inferences in observational studies.

# Appendix

## A 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 : Consider a full rank set of covariates, $X = (\mathbf{1}, X_1, ..., X_P)$ and a treatment variable, $T$, where $T$ is an $n \times 1$ vector, $\mathbf{1}$ is an $n \times 1$ vector of ones, and $X_p$ is an $n \times 1$ vector covariate for each $p$. Let $\widehat{T} = (X^\top X)^{-1} X^\top T$ be the linear predictor of $T$. The partial correlation of $T$ with each $X_p$ is zero given $\widehat{T}$, i.e., the second component of $(\widetilde{X}^\top \widetilde{X})^{-1} \widetilde{X}^\top T$ is zero, where $\widetilde{X} = (\mathbf{1}, X_p, \widehat{T})$.

PROOF : If we substitute $\widehat{T} = (X^\top X)^{-1} X^\top T$ into $(\widetilde{X}^\top \widetilde{X})^{-1} \widetilde{X}^\top T$ and use the identities $\mathbf{1}^\top \widehat{T} = \mathbf{1}^\top T$, $X_p^\top \widehat{T} = X_p T$, and $\widehat{T}^\top \widehat{T} = \widehat{T}^\top T$, the result follows from simple algebraic manipulations. ∎

## 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.

Angrist, J. D. and Krueger, A. B. (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.

Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business and 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, Jr., R. B. 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.

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 and Economic Statistics* **1**, 115–126.

Efron, B. and Feldman, D. (1991). Compliance as an explanatory variable in clinical trials (with discussions). *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.

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). *Research Designs: Inspired by the Work of Donald Campbell (eds. L. Bickman)*, chap. The Design of the New York School Choice Scholarship Program Evaluation, 155–180. Sage, Thousand Oaks, CA.

Hirano, K., Imbens, G., and Ridder, G. (2002). Efficient estimation of average treatment effects using the estimated propensity score. *forthcoming in Econometrica* .

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

Imai, K. (2003). Do get-out-the-vote calls reduce turnout? the importance of statistical methods for field experiments. *Revised for American Political Science Review* .

Imai, K. and van Dyk, D. A. (2003). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Revised for Journal of Econometrics* .

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,* to appear.

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 and Economic Statistics* **19**, 358–364.

Larsen, M. D. (1999). An analysis of survey data on smoking using propensity scores. *Sankhya, Series B, Indian Journal of Statistics* **61**, 91–105.

Lechner, M. (1999). Earnings and employment effects of continuous off-the-job training in East Germany after unification. *Journal of Business and Economic Statistics* **17**, 74–90.

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 (Second edition)*. Chapman & Hall, London.

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 Bickel, P. J. and Kwon, J. *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. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association* **79**, 565–574.

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, Series B, Methodological* **45**, 212–218.

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

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

Rosenbaum, P. R. and Rubin, D. B. (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 and Economic Statistics* **13**, 217–224.

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

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

Rubin, D. B. (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.

Rubin, D. B. (2000). *Statistical Science in the Courtroom (J. L. Gastwirth eds.)*, chap. Statistical Issues in the Estimation of the Causal Effects of Smoking Due to the Conduct of the Tobacco Industry, 321–351. Springer, New York.

Rubin, D. B. (2001). Estimating the causal effect of smoking. *Statistics in Medicine* **20**, 1395–1414.

Rubin, D. B. and Thomas, N. (1992). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika* **79**, 797–809.

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

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65**, 557–586.

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