Propensity Score Reweighting and Changes in Wage Distributions

John DiNardo
University of Michigan and NBER *

July 2002
Preliminary and Incomplete – Comments Welcome

Abstract

I provide a simple introduction to the use of propensity score reweighting to assess the effect of changes in covariates on the distribution of an outcome (such as wages). I relate this to the literature on estimating "average treatment effects" and Blinder/Oaxaca decompositions as well as discuss some of the limitations and uses of reweighting.

I. Introduction

A large literature in labor economics has been interested in the question: what would the distribution (mean/median/etc) of wages (or other outcome) look like if the covariates were different than they actually are. For instance, what would the (mean) wage for women be if women had the same distribution of human capital characteristics as men? (I.e. Blinder (1973) and Oaxaca (1973))

In this short, non-technical introduction I discuss the use of the *propensity score* as weights as an alternative to regression techniques, although such weights can also be used in other regression-based decomposition techniques such as those suggested by Juhn, Murphy and Pierce (1993). [1] In particular, I wish to describe the method proposed by DiNardo, Fortin and Lemieux (1996) (henceforth DFL) to analyze the effect of covariates on the distribution of wages and relate it to the literature on the use of the propensity score to analyze average treatment effects (Rosenbaum and Rubin (1983) Hirano, Imbens and Ridder (2000)) in part to highlight the limits of such methods.

*Prepared for the Latin American Meeting of the Econometrica Society

[1] See Lemieux (2002) for a discussion.

II. Three Types of Decompositions

The class of problems we will be concerned with will involved two "states" (s, t) and one continuous outcome, y, which will be a function of a set of exogenous covariates. As will be discussed in great detail below, I will consider three procedures:

1. The Blinder/Oaxaca method which involves running two separate regressions of y on x – one for each group, s, and t. Counterfactuals are constructed by using the coefficients from these two regressions and applying them to the x variables from either group s or group t depending on the question of interest.

2. The DFL method which involves using functions of the estimated propensity score[2] as weights in weighted kernel density estimation.

3. The use of weighted means to generate estimates of "average treatment effects" as suggested by Rosenbaum and Rubin (1983).

The procedures differ in one of two ways:

1. Are "counterfactual" weights used? (as in DFL and Rosenbaum and Rubin (1983))

2. Is the focus of interest a particular mean (as in Rosenbaum and Rubin (1983) and in the Blinder/Oaxaca procedure) or the entire distribution (as in DFL)?

The following table shows some illustrative cases from the empirical literature.

[2]As will be discussed below, the estimated propensity score is merely the predicted value from a binary dependent variable model where the dependent variable is 0 or 1 as a given observation is in group s or group t.

Some Illustrative Examples

Question:	Focus	Paper	s	t
How would the wage distribution at time t have looked if the distribution of covariates was as it was in time s	Distribution	DFL (1996),	1973	1993
What would the wages of Women be if they had the human capital characteristics of Men?	Mean	Blinder (1973) Oaxaca (1973)	Men	Women
What would the wealth of Blacks be if they had the demographic characteristics of whites?	Mean and Distribution	Barsky, Bound, Charles, Lupton JASA (Forthcoming)	Whites	Blacks
What is the Average Treatment Effect (ATE)	Mean	Rosenbaum and Rubin (1983)	Treatment	Control
What was the effect of changes in household structure and changes in returns on household inequality?	Distribution	Hyslop and Maré (2000)	1983-1986	1995-98
How did the transition to services effect wage inequality?	Distribution	Daly and Valletta (2000)	Manufacturing	Services

In choosing these examples, I have deliberately restricted myself to less "sophisticated" structures – some of which use some sort of weighting procedure as an intermediate step, others which avoid weighting altogether and provide very different approaches to the same sorts of questions: some examples include, François Bourguignon, Francisco Ferreira and Leite (2002), Donald, Green and Paarsch (2000), Lemieux (2002), Manacorda (1999), Teulings (2000), Lee (1999), Machado and Mata (2001). While I will not discuss the advantages and disadvantages of weighting to these methods, most of the address some aspect of the limitations of weighting, a subject I will address.

III. Blinder/Oaxaca Decomposition as a reweighting technique

The simplest case of a Blinder/Oaxaca "decomposition" involves a continuous outcome (say wages) and a single categorical variable. A question which this approach is designed to answer is:

What would average wages look like today (2002), if the distribution of characteristics (say schooling, experience, etc.) were held to their level in 1979? For now, assume random sampling, although sample weights can be dealt with easily.

We will focus on the case of a single covariate which can take on any one of $k \in (1 \dots k)$ values. This is slightly more general than might appear at first: if there are multiple categorical covariates, one should think of the list of categories for the single variable case as referring to all permutations of the multiple categorical variables. For example, if there are two genders and two education categories (high and low), the single variable takes on one of 4 values – male and low education, male and high education, female and low education, female and high education.

The Blinder/Oaxaca decomposition starts from a simple regression of y_j on a set of dummy variables for the different categories (where $D_{ij} = 1 \quad \text{if} \quad x_j = i \quad \text{and} \quad 0 \quad \text{otherwise}$):

$$\hat{y}_{ij} = \sum_{i}^{k} \hat{\beta}_i^{02} D_{ij} \tag{1}$$

where $\hat{y}_{ij}$ is the "predicted" wage from a regression of wages in 2002 on a complete set of dummy variables and where $\hat{\beta}$ are the associated OLS coefficients.

In this (trivial) case, the average wage that would have prevailed given the 2002 wage structure and the distribution of x that prevailed in 2002 is merely the average predicted wage for all observations in 2002. Moreover, this average wage can be expressed as the weighted sum of the k group averages, where the weights are the proportions of type i in the 2002 population:

$$\theta_i^{02} \equiv \frac{1}{N_{02}} \sum_{j \in 02} D_{ij}$$

Observe that $\hat{\beta}_i$ is merely the average wage of workers in 2002 with characteristic i. Therefore this can also be written as a weighted sum of group averages $\hat{\beta}_i^{02} \equiv \overline{y}_i^{02}$ where the weights θ_i^{02} are merely the 2002 sample fraction of observations in group i:

$$\overline{y}_{02}^{02} = \frac{1}{N_{02}} \sum_{j \in 02} \sum_{i}^{k} \hat{\beta}_i^{02} D_{ij}$$

$$
\begin{aligned}
&= \sum_{i}^{k} \hat{\beta}_i^{02} \frac{1}{N_{02}} \sum_{i=1}^{N_{02}} D_{ij} \\
&= \sum_{i}^{k} \hat{\beta}_i^{02} \theta_i^{02} \\
&= \sum_{i}^{k} \overline{y}_i^{02} \theta_i^{02} \qquad (2) \\
&= \theta_1^{02}\overline{y}_1^{02} + \theta_2^{02}\overline{y}_2^{02} + \theta_3^{02}\overline{y}_3^{02} + \ldots + \theta_k^{02}\overline{y}_k^{02} \\
&= \overline{Y}_{02} \qquad (3)
\end{aligned}
$$

where N_{02} is the number of observations in 2002, and $\overline{Y}_{02}$ is the actual sample average.

Likewise, we can compute the average wage that would have prevailed in 1979 given the distribution of x that prevailed in 1979 by taking the average predicted wage – using the 1979 coefficients for the 1979 observations. Noting that $\frac{1}{N_{79}} \sum_{j \in 79} D_{ij} \equiv \theta_i^{79}$, this of course, yields the actual 1979 wage:

$$
\begin{aligned}
\overline{y}_{79}^{79} &= \frac{1}{N_{79}} \sum_{j \in 79} \sum_{i}^{k} \hat{\beta}_i^{79} D_{ij} \\
&= \sum_{i=1}^{k} \hat{\beta}_i^{79} \frac{1}{N_{79}} \sum_{j=1}^{N_{79}} D_{ij} \\
&= \sum_{i}^{k} \overline{y}_i^{79} \theta_i^{79} \qquad (4) \\
&= \theta_1^{79}\overline{y}_1^{79} + \theta_2^{79}\overline{y}_2^{79} + \theta_3^{79}\overline{y}_3^{79} + \ldots + \theta_k^{79}\overline{y}_k^{79} \\
&\equiv \overline{Y}_{79} \qquad (5)
\end{aligned}
$$

The object of the Blinder/Oaxaca method is to compute the counterfactual "the average wage that would have prevailed given the distribution of characteristics observed in 1979." This is accomplished simply by plugging 1979 sample of x variables, computing a "predicted counterfactual" using the 2002 coefficients, and then taking the average across all the observations in the 1979 sample. In our simple illustration, note that these predicted counterfactual wages are merely the mean wages in 2002 for our k groups. In notation, the average wage that would have prevailed if the distribution of x across individuals was as it was in 1979, but with the 2002 wage structure:

$$
\overline{y}_{79}^{02} = \frac{1}{N_{79}} \sum_{j \in 79} \sum_{i}^{k} \hat{\beta}_i^{02} D_{ij} \qquad (6)
$$

$$
\begin{aligned}
&= \sum_{i}^{k} \hat{\beta}_i^{02} \frac{1}{N_{79}} \sum_{j \in 79} D_{ij} && (7)\\
&= \sum_{i \in \Omega} \theta_i^{79} \hat{\beta}_i^{02} && (8)\\
&= \theta_1^{79}\overline{y}_1^{02} + \theta_2^{79}\overline{y}_2^{02} + \theta_3^{79}\overline{y}_3^{02} + \ldots + \theta_k^{79}\overline{y}_k^{02}\\
&\equiv \sum_{i \in \Omega} \theta_i^{79}\overline{y}_i^{02} && (9)
\end{aligned}
$$

Observe that this is identicial to the actual 2002 mean wage, with θ_i^{79} replacing θ_i^{02}.

The effect of the changing "wage structure" (or to use a more popular term, the "price" of ability) on average wages is merely the difference between the actual mean wage and the counterfactual mean wage.

$$\text{Effect} = \overline{y}_{02}^{02} - \overline{y}_{02}^{79} \tag{10}$$

For my purposes, a useful observation about the Blinder/Oaxaca method is that in this simple case, it can be viewed as a "reweighting" method (and indeed in this case is identical to the method proposed by DFL which I discuss below.)

We have already noted that the actual mean wage – y_{02}^{02} is a weighted sum of the 2002 group means, where the weights are merely the proportion of each group in the 2002 sample. The key observation is that the Blinder/Oaxaca counterfactual wage – the 2002 wage structure and the 1979 x characteristics is also a weighted mean of the individual wage observations *in the 2002 data.*

Summarizing the above developments:

$$
\begin{aligned}
\text{"Counterfactual"} \quad \overline{y}_{79}^{02} &= \frac{1}{N^{79}} \sum_{j \in 79} \sum_{i}^{k} \hat{\beta}_i^{02} D_{ij} && (11)\\
&= \sum_{i}^{k} \theta_i^{79}\overline{y}_i^{02} && (12)\\
\text{"Actual"} \quad \overline{y}_{02}^{02} &= \frac{1}{N^{02}} \sum_{j \in 02} \sum_{i}^{k} \hat{\beta}_i^{02} D_{ij} && (13)\\
&= \sum_{i}^{k} \theta_i^{02}\overline{y}_i^{02} && (14)
\end{aligned}
$$

Inspection of the two equations above suggests that it will be useful to define a "weight" which is the ratio of the fraction of type i in the two time periods:

$$\omega_j = \frac{\theta_i^{79}}{\theta_i^{02}} \quad \text{if} \quad x_j = i \tag{15}$$

Indeed, weighting each observation in the 2002 sample by this weight yields the same counterfactual wage as in equation (9):

$$\begin{aligned}
\overline{y}_{79}^{02} &\equiv \frac{1}{N_{02}} \sum_{j \in 02} \omega_j y_j && (16)\\
&= \frac{1}{N_{02}} \sum_{j \in 02} \sum_{i}^{k} y_j^{02} D_{ij} \omega_j && (17)\\
&= \frac{1}{N_{02}} \sum_{i}^{k} \overline{y}_i^{02} \sum_{j \in 02} D_{ij} \omega_j && (18)\\
&= \sum_{i}^{k} \frac{\theta_i^{79}}{\theta_i^{02}} \overline{y}_i^{02} \left(\frac{1}{N_{02}} \sum_{j \in 02} D_{ij} \right) && (19)\\
&= \sum_{i}^{k} \left(\frac{\theta_i^{79}}{\theta_i^{02}} \right) \overline{y}_i^{02} \theta_i^{02} && (20)\\
&= \sum_{i}^{k} \theta_i^{79} \overline{y}_i^{02} && (21)
\end{aligned}$$

In this simple case, inspection was enough to suggest the correct weighting factor, although the appropriate weighting factor can be easily derived and we will do so below.

IV. The Use of Weights to Generate Counterfactual distributions

While the Blinder/Oaxaca is suitable for many applications, it is less well suited to the case of examining either the entire distribution of wages or moments other than the mean. After briefly defining a kernel density estimator, I describe the estimator used by DiNardo et al. (1996) to generate counterfactual distributions.

The kernel density case is merely a histogram of sorts where the bins are not mutually exclusive and one takes a weighted (instead of unweighted) average of points in a bin.

To introduce the notion of weights, consider the usual case where each observation has a known weight (usually the inverse of the sampling probability) w_i consider the definition of a kernel density estimator at the point y_0.

$$\text{pdf}(y_0) = \sum_{j=1}^{N} w_i \frac{1}{Nh} K\left(\frac{y_i - y_0}{h}\right) \quad (22)$$

where K is some kernel. The choice of kernel – a simple weighting function – is usually of little consequence. In the simple histogram, for example, this function is a constant for each

observation in a given bin and zero for observations outside the bin. For various reasons, including efficiency and the need for smooth derivatives, a variety of other functions are generally used. Many of these have the property that the weight an observation is given is a declining function of its distance from the center of the bin. The use of sample weights, (normalized so that $\sum_{j=1}^{N} \omega_j = 1$) is a straightforward extension of a kernel density estimate.

The key parameter in the density estimate, like in the histogram, is the "bandwidth" (or bin width in the case of the histogram). The wider the bandwidth, the smoother the density estimate. Indeed, there is a trade-off – the variance of the estimator declines with the bandwidth but the bias increases. In words, since it is easier to smooth with the eye than to unsmooth for many purposes it is better to err on the side of too small a bandwidth.

A Weighting with the propensity score

The DFL procedure can be derived fairly easily from the definition of conditional probability. Consider the two actual distributions:

$$\int f^{79}(y)dy \equiv \int f^{79}(y|x)h(x|t=79)dx \tag{23}$$

$$\int f^{02}(y)dy \equiv \int f^{02}(y|x)h(x|t=02)dx \tag{24}$$

Now consider the counterfactual distribution "the distribution of wages in 2002 if the distribution of x was as in 1979 and compare it to the actual distribution of wages in 2002:

$$\text{Counterfactual} \quad f_{79}^{02}(y) \equiv \int f^{02}(y|x)h(x|t=79)dx \tag{25}$$

$$\text{Actual} \quad f_{02}^{02}(y) \equiv \int f^{02}(y|x)h(x|t=02)dx \tag{26}$$

The important observation is that the counterfactual described differs from its actual counterpart in what set of x variables are to be "integrated over." While this is easily computed in this case, in general $h(x)$ will have several explanatory variables and integrating over several, possibly hundreds of covariates would be impossible.

As in the Blinder/Oaxaca case, the key will to be define a weight ω such that:

$$\int f^{02}(y|x)h(x|t=79)dx = \int \omega f^{02}(y|x)h(x|t=02)dx$$

If we can find such a weight, we have transformed a potentially impossible problem – integrating over many covariates – with a simple one – weighting the 2002 distribution. To do so, it will help to consider pooling the 1979 and 2002 data and observe that by definition:

$$h(x_j = x_0) = \frac{h(x_j|t = 79)P_{79}}{P(t = 79|x_j = x_0)} \tag{27}$$

$$h(x_j = x_0) = \frac{h(x_j|t = 02)P_{02}}{P(t = 02|x_j = x_0)} \tag{28}$$

where $P_{79} = P(\text{Observation is from 1979})$ and

$$\rho^{02}(x) \equiv P(t = 02|x_j = x_0)$$

$$\rho^{79}(x) \equiv P(t = 79|x_j = x_0)$$

are the propensity scores associated with being either in 2002 or 1979. The utility of these expressions comes from the fact that it is much more difficult to integrate over a term like $h(x_j|t = 02)$ which requires multi-dimensional integration than a term like $P(t = 79|x_j = x_0)$ which looks like a standard binary dependent variable model like the logit or probit. This latter term is called the propensity score. It is a number between 0 and 1 and in this case can be interpreted as the probability I will given observation will be from 2002 (1979) given a set of characteristics. To use the language of experiments, the "treatment" is exposure to the 2002 (1979) wage structure and the propensity score is the probability I have been exposed given my characteristics. For instance, if the characteristic of interest is "has a college degree", and this has been rising over time:

$$\rho^{02}(\text{Has college degree}) > \rho^{02}(\text{Does not have college degree})$$

One does not observe the propensity score in general, but fortunately, an estimate of $\rho^{79}(x)$, for example, can easily be generated by:

1. Pooling the 2002 and 1979 data.

2. Define a binary variable T such that $T = 1$ if the observation is from the 1979 sample and 0 otherwise.

3. Run a logit or other estimator using T as the dependent variable with a flexible specification of the relevant *exogenous* covariates. The predicted probability from the logit is an estimate of the propensity score.

As we noted above, observe that the the actual and counterfactual distributions differ only in the term $h(x|t = 79 \text{ or } 02)$ Simple manipulations with equations (27) and (28) reveal that the weight is a simple function of the propensity score and two constants:

$$
\begin{aligned}
\text{Counterfactual} \quad f_{79}^{02}(y) &\equiv \int f^{02}(y|x)h(x|t=79)dx \\
&= \int \left(\frac{\rho^{79}(x)}{1-\rho^{79}(x)}\right)\left(\frac{P_{02}}{P_{79}}\right) f^{02}(y|x)h(x|t=79)dx \qquad (29) \\
&= \int \omega f^{02}(y|x)h(x|t=02)dx \qquad (30)
\end{aligned}
$$

In the expression above P_{79} (P_{02}) are merely the fraction of the pooled sample that comes from 1979 (2002) and does not vary by observation. This term can essentially be ignored since most packages appropriate renormalize the weights to sum to 1.

To apply this to actual data with known sampling weights w_i, (where $\sum w_i = 1$) DiNardo et al. (1996) proposed going from the expression in equation (30) directly to density estimation by replacing ρ^{79} with an estimate of $\widehat{\rho}^{79}$ from a logit or other binary dependent model to construct $\widehat{\omega}_i$ and the product of this and the usual sampling weights (normalized so that the sum of the weights is equal to 1) is used in the density estimation:

$$
\begin{aligned}
\text{pdf}(y_{02}^{02}) &= \sum_{j=1}^{N} w_i \frac{1}{Nh} K(\frac{y_i - y_0}{h}) \\
\text{pdf}(y_{79}^{02}) &= \sum_{j=1}^{N} \widehat{\omega}_i w_i \frac{1}{Nh} K(\frac{y_i - y_0}{h})
\end{aligned} \qquad (31)
$$

NB: It is not necessary to actual do any kernel density estimation if only simple sample statistics are necessary. For example, to compute the counterfactual mean

$$
\begin{aligned}
\overline{y}_{02}^{02} &= \sum_{j \in 02} w_j y_j \\
\overline{y}_{79}^{02} &= \sum_{j \in 02} \widehat{\omega}_j w_j y_j
\end{aligned} \qquad (32)
$$

where the weights are suitably normalized as in the density case.

It remains to be shown that this is numerically identical to the Blinder/Oaxaca counterfactual when there is only one categorical covariate. To do so, without loss of generality let the sample weight be the same for every observation, $\frac{1}{N_{02}}$ so that we merely have to show that :

$$\left(\frac{\widehat{\rho}^{79}(x)}{1-\widehat{\rho}^{79}(x)}\right)\left(\frac{P_{02}}{P_{79}}\right)=\frac{\theta_i^{79}}{\theta_i^{02}} \tag{33}$$

When we have the case of one categorical covariate[3], any appropriate estimator of the binary dependent variable model will generate the numerically equivalent weight.[4] To see this, consider the 1979 propensity score – the predicted probability that an observation from the pooled 1979 and 2002 sample comes from 1979. I.e., the predicted value associated with an observation x_i such that $x_i = j$:

$$\begin{aligned}
\widehat{\rho}^{79}(x_i|x_i=j) &= \widehat{Pr}(\text{observation } i \text{ from 79 sample}|x_i=j) \\
&= \frac{\widehat{Pr}(\text{observation } i \text{ from 79 sample} \wedge x_i=j)}{\widehat{Pr}(x_i=j)} \\
&= \frac{\frac{\sum_{i\in 79} D_{ij}}{N_{79}+N_{02}}}{\frac{\sum_{i\in 02} D_{ij}+\sum_{i\in 79} D_{ij}}{N_{79}+N_{02}}} \\
&= \frac{P_{79}\theta_j^{79}}{P_{79}\theta_j^{79}+P_{02}\theta_j^{02}}
\end{aligned} \tag{34}$$

It is now straightforward to observe equation 33 holds and where $(1-\widehat{\rho}^{79}(x)) \equiv \widehat{\rho}^{02}(x)$ so that

$$\widehat{\rho}^{02} = P_{02}\frac{P_{02}\theta_j^{02}}{P_{79}\theta_j^{79}+P_{02}\theta_j^{02}}$$

So:

$$\begin{aligned}
\frac{\widehat{\rho}^{79}}{\widehat{\rho}^{02}} &= \frac{\frac{P_{79}\theta_j^{79}}{P_{79}\theta_j^{79}+P_{02}\theta_j^{02}}}{\frac{P_{02}\theta_j^{02}}{P_{79}\theta_j^{79}+P_{02}\theta_j^{02}}} \\
&= \frac{P_{79}\theta_j^{79}}{P_{02}\theta_j^{79}}
\end{aligned} \tag{35}$$

[3]We are also assuming that their are no empty cells – i.e. that their are individuals of type j in both samples. This can be a problem in some contexts. Barsky, Bound, Charles and Lupton (2002) for example, in their analysis of black/white wealth distributions observe that while one can generally find whites to "match" with any black in a sample, the reverse is not the case.

[4]For a proof in a different context see Davidson and MacKinnon (1993), page 234, for example.

and and the equality in equation (33) follows.

In the case when the covariates can be described by a set of dummy variables, the Blinder/Oaxaca and DFL methods are numerically equivalent. When this is not the same, the two methods will not be numerically equivalent, but should be "close" if the specification of the binary choice model is sufficiently flexible.

B Reweighting for Estimates of Average Treatment Effects

Having now shown that the Blinder/Oaxaca estimate of the counterfactual mean is numerically equivalent to using "DFL" weights in the simple case it may be useful to relate DFL to the literature that uses the propensity score to compute "average treatment effects. " Consider the evaluation of a randomized experiment where $T = 1$ if the person gets the treatment and zero otherwise. Again, let the outcome y depend on some covariates so we have the distribution of outcomes in the treatment group and the distribution of outcomes in the control group ($T = 0$).

$$\int f^{T=1}(y)dy = \int f^{1}(y|x)h(x|T=1)dx \tag{36}$$

$$\int f^{T=0}(y)dy = \int f^{0}(y|x)h(x|T=0)dx \tag{37}$$

In that literature, the focus of much attention has been on average treatment effect which is merely the treatment effect averaged over all "types" which is given by x

$$\text{ATE} = \int yf^{1}(y|x)h(x)dx - \int yf^{0}(y|x)h(x)dx \tag{38}$$

In general, the simple difference between the treatment and control group means will not be the average treatment effect. Denote this simple difference in means as the "usual":

$$\text{Usual} = \int yf^{1}(y|x)h(x|T=1)dx - \int yf^{0}(y|x)h(x|T=0)dx \tag{39}$$

The key is that, in general, $h(x|T = 1) \neq h(x)$ and likewise for the control group so that the estimates in equation 38 and equation 39 will generally not be equal.

From the perspective of DiNardo et al. (1996) ATE is actually the difference between *two* counterfactuals:

1. The average outcome using the "structure of outcomes under treatment" *as if* the treatment had been given to both the treatment and the controls (i.e. the general population)

2. The average outcome using the "structure of outcomes under control" *as if* the "control had been given" to both the treatment and the controls (i.e. the general population)

Using the same trick as before, we can reweight the usual estimator so that both means are averaged over the entire population and take the difference. Using the definition of conditional probability and expressions like equations (27) and (28) we see that each term on the right hand side of equation (38) are merely weighted versions of the actual means of the treated and control groups:

$$\begin{aligned}
\text{Actual Treatment Mean} &= \int y f^1(y|x)h(x|T=1)dx && (40)\\
\text{"CF" Treatment Mean} &= \int y f^1(y|x)h(x)dx && (41)\\
&= \int \left(\frac{P_1}{\rho(x)}\right) f^1(y|x)h(x|T=1)dx && (42)\\
\text{Actual Control Mean} &= \int y f^0(y|x)h(x|T=0)dx && (43)\\
\text{"CF" Control Mean} &= \int y f^0(y|x)h(x)dx && (44)\\
&= \int \left(\frac{P_0}{1-\rho(x)}\right) y f^0(y|x)h(x|T=0)dx && (45)\\
ATE &= \int \left(\frac{P_1}{\rho(x)}\right) f^1(y|x)h(x|T=1)dx\\
&- \int \left(\frac{P_0}{1-\rho(x)}\right) y f^0(y|x)h(x|T=0)dx && (46)\\
\widehat{\text{ATE}} &= \sum_{i\in T} \widehat{\omega}_i^T y_i - \sum_{i\in C} \widehat{\omega}_i^C y_i && (47)
\end{aligned}$$

where P_1 is the fraction of treatment observations, P_0 is the fraction of control observations and where

$$\widehat{\omega}_i^T = \frac{P_1}{\widehat{\rho}^T(x)}$$

and

$$\widehat{\omega}_i^C = \frac{P_0}{1-\widehat{\rho}^T(x)}$$

where $\widehat{\rho}^T x$ is the predicted probability from a binary dependent variable model where the dependent variable is equal to one when the observation is from the treatment group and

zero otherwise. If the $\widehat{\omega}$ do not sum to one over their respective samples, they should be normalized so that they do.

Indeed the weighting estimator proposed by Rosenbaum and Rubin (1983) and Hirano et al. (2000) is this estimator.

C Effect of the Treatment on the Treated (TOT)

Likewise the same type of estimator can be used to estimate other outcomes. One might be interested, for example, in the effect of the treatment on the treated:

$$\begin{aligned}
\text{Actual Treatment Mean} &= \int y f^1(y|x)h(x|T=1)dx && (48)\\
\text{Actual Control Mean} &= \int y f^0(y|x)h(x|T=0)dx && (49)\\
\text{CF Control Mean} &= \int y f^0(y|x)h(x|T=1)dx && (50)\\
\text{TOT} &= \int y f^1(y|x)h(x|T=1)dx - \\
&\quad \int y f^0(y|x)h(x|T=1)dx && (51)\\
&= \int y f^1(y|x)h(x|T=1)dx - \\
\int \omega' y f^0(y|x)h(x|T=0)dx && (52)
\end{aligned}$$

where

$$\omega' = \frac{\rho(x)}{1-\rho(x)}\frac{P_0}{P_1} \tag{53}$$

where as before the term $\frac{P_0}{P_1}$ is merely a constant and the weight can be renormalized to sum to one.

That is, the effect of the treatment on the treated is merely the treatment mean less the reweighted control group mean.

D Other Uses of Propensity Score Weighting

While the results here are not new (Hirano et al. (2000) for example show that these weighted versions are more efficient that "regression – on – covariates") one observation that has not been made is that the propensity score weighting techniques can be used to

learn about other aspects of the treatment besides its effect on the mean. Since the use of weights allows for the estimation of the entire distribution, any statistic can be computed.

For example, one could compute a counterfactual variance that would correspond to the variance if the distribution of x were as it were in the entire sample:

$$\sum \widehat{\omega}_i (y_i - \tilde{y})^2$$

where $\tilde{y} = \sum \widehat{\omega}_i y$ and the form of $\widehat{\omega}$ would be the normalized version of $\frac{1}{\widehat{\rho}_i}$ and the effect of the treatment on the variance – the "variance" treatment effect – would be given by the difference between this and the appropriate "counterfactual" control variance (i.e. the variance among the controls that would have obtained if the distribution of x in the control group were the same in the pooled distribution of x as above.)

Lemieux (2002), for example, observes that this can be applied to the type of decomposition proposed by Juhn et al. (1993), where

$$y = X\hat{\beta} + e \tag{54}$$

where[5], the variance of y is decomposed into a part due to the observables $X\hat{\beta}$ and a part due to the unobservables e. In this framework, Juhn et al. (1993) propose to treat increases in the variance of e over time as evidence that the "price" of unobservables has risen. Lemieux (2002) observes that under the assumptions of this framework, however, the variance of e will in general depend on the distribution of X, so that merely comparing the variance of unobservables over two time periods is not sufficient to judge whether the price has risen over time – one has to hold fixed the distribution of X in the population. Lemieux (2002) proposes using propensity score weighting as a simple expedient to accomplish this goal.

V. Implications for the use of Weighting

Several limitations of weighting become immediately apparent when one views applications such as DiNardo et al. (1996) as either extensions of propensity score weighting in Rosenbaum and Rubin (1983) and Hirano et al. (2000) or as variant of the Blinder/Oaxaca decomposition method. Indeed, the both literatures have identified the substantial limitations of such approaches and I mention only a few here.

1. Implicit in our discussion has been the following model:

$$y_i = \beta_i^0(x_i) + \beta_i^1(x_i)T_i + \epsilon_i \tag{55}$$

[5]Ignoring the fact that the residual e contains sampling error in addition to the unobservable error term ϵ

where the heterogeneity in the treatment effect is reflected in the fact that β is now a random variable and quite importantly ϵ_i is assumed to be independent of assignment to the treatment. In particular, assignment to treatment is assumed to depend only on the *observables*. However, if individuals "select in" to treatment on the basis of factors not observed by the econometrician, an important literature, pioneered by James Heckman shows that the proposed estimators can be seriously biased. Intuitively, propensity score reweighting is about averaging the difference between treated and untreated individuals who are identical in all *observable* respects save the treatment across individuals so paired. However, if we observe two individuals who are identical in a long list of observables, save the fact that one receives the treatment and the other does not, in many contexts it is not appropriate to assume that the individuals are identical in ways not observed by the econometrician.

Using similar notation as above, the Heckman selection framework can be written

$$y_i = \beta_i^0(x_i, \epsilon_i) + \beta_i^1(x_i, \epsilon_i)T_i + \epsilon_i \tag{56}$$

It is therefore clear that this propensity score reweighting is merely a special case of the Heckman selection framework. Moreover, it is not clear that "average treatment effects" are always of interest. For some recent work, see Heckmand and Vytlacil (2001) and Heckman and Vytlacil (2001) and Heckman, Tobias and Vytlacil (2001).[6]

2. In principle, if the propensity score takes on small finite number, and one is interested only in averages these reweighting exercises could be accomplished by regression or matching techniques. One expedient, for example would involve running separate regressions for each value of the propensity score. While Hirano et al. (2000) observe that weighting is (asymptotically) more efficient regression or matching, convenience seems more important.

3. A related point is that reweighting methods are "methods of ignorance" in the sense that we are not being explicit about why treatment effects vary across individuals. In an analysis of union wage effects, Card (1992), for example, divides the sample into 5 groups, not on the basis of a propensity score, but on a predicted wage. Since are *a priori* grounds for believing that the effect of unionization depends in part on what wage you would have received in the non–union sector.

4. For propensity score weighting to be appropriate the x variables have to be exogenous. This rules out any sort of endogeneity or interactions. François Bourguignon et al. (2002) and Teulings (2000) are two examples that address these issues.

[6]For further discussion of the evaluation problem in this more general framework see Heckman, LaLonde and Smith (1998b), Heckman and Hotz (1989), Heckman and Robb (1984), Heckman, Ichimura, Smith and Todd (1998a), and the references cited therein and above.

5. Extremely low (or high) values of the propensity score are a potential problem. Intuitively, if the propensity score for having received the treatment is very small, this means that there are none (or few) treatment observations that "look like" the corresponding "control" observation.

6. A related point is that the weights are often of the form $\frac{1}{\rho(x)}$ or $\frac{1}{1-\rho(x)}$. In this case, small errors in estimating $\rho(x)$ can produce potentially large errors in the weights. Since the weight is a nuisance parameter from the viewpoint of estimating a density or a specific moment of the distribution, this is not a straightforward problem.

7. In the context of density estimation, very little work has been done on estimating standard errors. In part, this is because of a problem in the literature on density estimation.

Given such negatives, is there anything good to say about propensity score reweighting methods?

1. It's easy.

2. It's usually an interesting "base" case.

3. It is often a helpful descriptive tool.

References

Barsky, Robert, John Bound, Kerwin Charles, and Joseph Lupton, "Accounting for the Black-White Wealth Gap: A Nonparametric Approach," *Journal of the American Statistical Association*, 2002, forthcoming.

Blinder, Alan S., "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, Fall 1973, *8*, 436–455.

Card, David, "The Effect of Unions on the Distribution of Wages: Redistribution or Relabelling?," NBER Working Paper 4195, National Bureau of Economic Research, Cambridge, MA. October 1992.

Daly, Mary C. and Robert G. Valleta, "Inequality and Poverty in the United States: The Effects of Changing Family Behavior and Rising Wage Dispersion," Working Paper in Applied Economic Theory 2000-06, Federal Reserve Bank of San Francisco, San Francisco, CA 2000.

Davidson, Russell and James G. MacKinnon, *Estimation and inference in econometrics*, New York: Oxford University Press, 1993.

DiNardo, John, Nicole Fortin, and Thomas Lemieux, "Labor Market Institutions and The Distribution of Wages, 1973-1993: A Semi-Parametric Approach," *Econometrica*, September 1996, *64* (5), 1001–1045.

Donald, Stephen G., David A. Green, and Harry A. Paarsch, "Differences in Wage Distributions between Canada and the United States: An Application of a Flexible Estimator of Distribution Functions in the Presence of Covariates," *Review of Economic Studies*, 2000, *67*, 609–633.

François Bourguignon, Francisco Ferreira, and Phillipe George Leite, *Beyond Oaxaca-Blinder: accounting for differences in household income distributions across countries* number PUC-Rio, TD #452. In 'Working Paper.' March 2002.

Heckman, James and Edward Vytlacil, "Structural Equations, Treatment Effects and Econometric Policy Evaluation," 2001.

___ and Joseph Hotz, "Alternative Methods for Evaluating the Impact of Training Programs," *Journal of the American Statistical Association*, 1989.

___ and R. Robb, "Alternative Methods for Evaluating the Impact of Interventions," in James Heckman and R. Singer, eds., *Longitudinal Analysis of Labor Market Data*, Cambridge University Press Cambridge 1984.

___ , H. Ichimura, J. Smith, and Petra Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 1998, *66*, 1017–1098.

Heckman, James J., Robert J. LaLonde, and James A. Smith, "The Economics and Econometrics of Active Labour Market Programmes," in "The Handbook of Labour Economics," Vol. III North–Holland Amsterdam 1998.

Heckman, James, Justin Tobias, and Edward Vytlacil, "Simple Estimators for Treatment Parameters in a Latent Variable Framework," 2001.

Heckmand, James and Edward Vytlacil, "Causal Parameters, Structural Equations, Treatment Effects, and Randomized Evaluations of Social Programs," 2001.

Hirano, Keisuke, Guido Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," NBER Technical Working Paper T0251, National Bureau of Economic Research, Cambridge, MA April 2000.

Hyslop, Dean and David Maré, "Understanding Changes in the Distribution of Household Incomes in New Zealand Between 1983-86 and 1995-98," Treasury Working Paper 1/21, New Zealand Department of the Treasury, Wellington, NZ 2001.

Juhn, Chinhui, Kevin M. Murphy, and Brooks Pierce, "Wage Inequality and the Rise in the Returns to Skill," *Journal of Political Economy*, June 1993, *101* (3), 410–442.

Lee, David S., "Wage inequality in the United States during the 1980s: Rising dispersion or falling minimum wage?," *Quarterly Journal of Economics*, August 1999, *114* (3), 977–1023.

Lemieux, Thomas, "Decomposing Changes in Wage Distributions: A Unified Approach," *Canadian Journal of Economics*, 2002, *forthcoming.*

Machado, José and José Mata, "Counterfactual Decompositions of Changes in Wage Distributions using Quantile Regression," *Empirical Economics*, 2001, *26*, 115–134.

Manacorda, Marco, "The Evolution of Earnings Inequality in Italy and The Escalator Clause," Working Paper 16, Center for Labor Economics, University of California – Berkeley April 1999.

Oaxaca, Ronald, "Male–Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 1973, *14*, 693–709.

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

Teulings, Coen, "Aggregation Bias in Elasticities of Substitution and the Minimum Wage Paradox," *International Economic Review*, 2000, *41*, 359–398.