Oaxaca-Blinder as a Reweighting Estimator

By Patrick Kline*

A large literature focuses on the use of propensity score methods as a semi-parametric alternative to regression for estimation of average treatments effects.[1] We show here that the classic regression based estimator of counterfactual means studied by Ronald Oaxaca (1973) and Alan Blinder (1973) constitutes a propensity score reweighting estimator based upon a linear model for the conditional odds of being treated – a functional form which emerges, for example, from an assignment model with a latent log-logistic error.[2]

As such it enjoys the status of a "doubly robust" estimator of counterfactuals as in Robins, Rotnitzky, and Zhao (1994) – estimation is consistent if *either* the propensity score assumption or the model for outcomes is correct. To illustrate the method, the Oaxaca-Blinder estimator is applied to LaLonde's (1986) study of the National Supported Work program where it is found to compare favorably with competing approaches.

I. The Oaxaca-Blinder Estimator

Consider a population of individuals falling into two groups indexed by $D_i \in \{0,1\}$. We will refer to observations with $D_i = 1$ as the treatment group and those with $D_i = 0$ as the controls. Let X_i be a $K \times 1$ vector of random covariates (which we assume includes an intercept) and Y_i some outcome of interest. We begin by indexing the potential outcomes associated with treatment as follows:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

where Y_i^1 is the outcome individual i would experience if treated and Y_i^0 the outcome that would obtain in the absence of treatment.

The Oaxaca-Blinder (O-B) approach is predicated upon a model for the potential outcomes of the form:

$$(1) \qquad Y_i^d = X_i' \beta^d + \varepsilon_i^d$$

$$(2) \qquad E\left[\varepsilon_i^d | X_i, D_i\right] = 0 \text{ for } d \in \{0,1\}$$

Hence, knowledge of $\left(\beta^1, \beta^0\right)$ is sufficient to compute counterfactual means for either group. Natural estimators of these parameters come from linear regression in the two populations indexed by D_i.

Suppose in particular that we are interested in the counterfactual mean outcomes the treatment group would have experienced in the absence of treatment, a quantity we denote as:

$$\mu_0^1 \equiv E\left[Y_i^0 | D_i = 1\right]$$

We assume throughout that $E\left[X_i X_i' | D_i = 0\right]$ is finite and invertible so that a regression among the controls identifies β^0. According to the model in (1) and (2):

$$\begin{aligned} \mu_0^1 &= E\left[X_i | D_i = 1\right]' \beta^0 \\ &= E\left[X_i | D_i = 1\right]' \\ &\quad \times E\left[X_i X_i' | D_i = 0\right]^{-1} E\left[X_i Y_i | D_i = 0\right] \\ &\equiv \delta^{OB} \end{aligned}$$

When each of the moments in δ^{OB} is replaced by its sample analogue one obtains the O-B estimate of the counterfactual mean, which, by standard arguments, can be shown to be consistent for the parameter of interest. This estimator may be particularly convenient in settings where K is large and few treated observations are available as estimation only requires that collinearity

* UC Berkeley, 508-1 Evans Hall #3880, Berkeley, CA 94720-3880, pkline@econ.berkeley.edu. I thank Josh Angrist, David Card, Justin McCrary, Bryan Graham, and Andres Santos for helpful comments.

[1] Imbens (2004) provides a review.

[2] Dinardo (2002) shows the equivalence of nonparametric Oaxaca-Blinder and propensity score methods in the special case of discrete covariates.

problems be absent among the controls.[3]

II. Reweighting Estimators

A popular alternative to regression based methods is to use propensity score weighted averages of outcomes as estimates of counterfactual means. This approach is typically motivated by the following conditional independence assumption:

$$(3) \qquad \left(Y_i^1, Y_i^0\right) \perp\!\!\!\perp D_i | X_i$$

This restriction, termed "unconfoundedness" by Rosenbaum and Rubin (1983), amounts to assuming that treatment status was assigned randomly conditional on covariates. Note that the parametric O-B model would satisfy this condition were we to strengthen the mean independence assumption (2) to encompass full conditional independence of the errors.[4] However (3) is usually considered less restrictive than the O-B assumptions since it is agnostic about the dependence of the potential outcomes on the covariates. It is instructive then to consider the population moments that identify μ_0^1 using only the nonparametric restrictions inherent in (3).

We must first make the following "common support" assumption ensuring identification:

$$(4) \qquad e\left(X_i\right) < 1$$

where $e\left(X_i\right) \equiv P\left(D_i = 1|X_i\right)$ is the propensity score. This condition, which guarantees that suitable controls can be found for every treated unit, allows us to derive the following well-known result justifying the use of propensity score reweighting estimators:

PROPOSITION 1: *If* (3) *and* (4) *hold then:*

$$(5) \qquad \mu_0^1 = E\left[\frac{e\left(X_i\right)}{\pi}\frac{1-D_i}{1-e\left(X_i\right)}Y_i\right] = E\left[w\left(X_i\right)Y_i|D_i = 0\right]$$

where $w\left(X_i\right) \equiv \frac{1-\pi}{\pi}\frac{e(X_i)}{1-e(X_i)}$ *and* $\pi \equiv P\left(D_i = 1\right)$.

PROOF:

$$\begin{aligned} E\left[w\left(X_i\right)Y_i|D_i = 0\right] &= E\left[w\left(X_i\right)Y_i^0|D_i = 0\right] \\ &= E\left[w\left(X_i\right)E\left[Y_i^0|X_i\right]|D_i = 0\right] \\ &= \int E\left[Y_i^0|X_i = x\right]w\left(x\right)dF_{X|D=0}\left(x\right) \\ &= \int E\left[Y_i^0|X_i = x\right]dF_{X|D=1}\left(x\right) \\ &= E\left[Y_i^0|D_i = 1\right] \end{aligned}$$

The second line follows from (3) and the fourth from the fact that by Bayes' rule $\frac{dF_{X|D=1}(x)}{dF_{X|D=0}(x)} = w\left(x\right)$.

Thus, a weighted average of the control outcomes, with weights proportional to the conditional odds of treatment, identifies the counterfactual mean of the treated population. A large literature considers using sample analogues of (5) for estimation of μ_0^1, where $e\left(X_i\right)$ is replaced by some parametric or nonparametric estimator.[5] A difficulty with such approaches often arises in settings with few treated observations where simple propensity score models may perfectly predict treatment even if (4) holds in the population. Even when prediction is not perfect, recent studies suggests propensity score estimators which assign disproportionate weight to a few observations often exhibit poor finite sample performance.[6]

III. Equivalence

Let us now return to the parametric O-B estimand δ^{OB}. That this quantity has an interpretation as a weighted average of the control outcomes is self-evident. The following proposition shows that these weights have a propensity score based interpretation given only the common support assumption (4).

[3] See Busso, Gregory, and Kline (2010) for a recent application

[4] This would be equivalent to assuming in addition to (2) that $E\left[g\left(\varepsilon_i^d\right)|X_i, D_i\right] = E\left[g\left(\varepsilon_i^d\right)|X_i\right]$ for any continuous function $g\left(.\right)$ vanishing outside a finite interval and for $d \in \{0,1\}$. See e.g. Theorem 1.17 in Chapter V of Feller (1966).

[5] See Dinardo, Fortin, and Lemieux (1996), Hirano, Imbens, and Ridder (2003), and Imbens (2004).

[6] See Kang and Schafer (2007), Robins, Sued, Lei-Gomez, and Rotnitzky (2007), and Huber, Lechner, and Wunsch (2010).

PROPOSITION 2: *If* (4) *holds then:*

$$\begin{aligned}\delta^{OB} &= E\left[\widetilde{w}\left(X_i\right)Y_i|D_i=0\right]\\ \widetilde{w}\left(X_i\right) &= X_i'E\left[X_iX_i'|D_i=0\right]^{-1}\\ &\quad\times E\left[X_i\frac{1-\pi}{\pi}\frac{e\left(X_i\right)}{1-e\left(X_i\right)}|D_i=0\right]\end{aligned}$$

PROOF:

Bayes' rule and (4) imply $E\left[X_i|D_i=1\right]=E\left[X_i\frac{1-\pi}{\pi}\frac{e(X_i)}{1-e(X_i)}|D_i=0\right]$. Plugging this into the definition of δ^{OB} yields the result.

Note that the O-B weights $\widetilde{w}\left(X_i\right)$ are simply the normalized projection of the true treatment odds $\frac{e(X_i)}{1-e(X_i)}$ onto the column space of X_i – i.e. they are the predicted values from an (infeasible) population regression of $w\left(X_i\right)$ on X_i. Hence, the O-B specification provides a minimum mean squared error approximation to the true nonparametric weights $w\left(X_i\right)$.

Of course, if the true odds of treatment are actually linear in X_i, then $\widetilde{w}\left(X_i\right)=w\left(X_i\right)$, and Proposition 1 implies the O-B estimand will identify μ_0^1 even if the model for the outcomes is misspecified provided that (3) and (4) hold. A linear model for the treatment odds arises naturally from an assignment model of the form:

$$D_i=1\left[X_i'\delta+v_i>0\right]$$

where $1\left[.\right]$ is an indicator for whether the condition in brackets is true and the assignment error v_i is an *iid* draw from a standardized log-logistic distribution with CDF $F\left(z\right)=\frac{z}{1+z}$.[7]

Conversely, if the model for the outcomes in (1) and (2) is correct, the O-B estimand will identify μ_0^1 even if the common support condition (4) fails and/or the implicit model for the propensity score is incorrect. Hence the estimator is "doubly robust" (Robins, Rotnitzky, and Zhao, 1994) as it identifies the parameter of interest under two independent sets of assumptions.

A Remark on Misspecification

The double robustness property offers little comfort to the applied econometrician who suspects any propensity score model, like any model for the conditional mean, to provide only a rough approximation to the data generating process. Note from Propositions 1 and 2 that the population bias in the O-B approximation may be written:

$$\mu_0^1-\delta^{OB}=E\left[\left(w\left(X_i\right)-\widetilde{w}\left(X_i\right)\right)Y_i|D_i=0\right]$$

Though the O-B weights may yield specification errors at particular values of X_i, those errors will only induce bias if they are correlated with outcomes in the control sample.[8] If, for instance, $\widetilde{w}\left(X_i\right)=w\left(X_i\right)+\xi_i$ where ξ_i is a random specification error obeying $E\left[\xi_iY_i|D_i=0\right]=0$ then the O-B estimator will retain consistency.

An important question then is whether, in the absence of prior knowledge of the propensity score, approximations ought to be sought with respect to the propensity score or the weights themselves.[9] The O-B approach follows the latter approach, conventional propensity score methods the former. Which approach removes more bias in a misspecified environment will depend on the specifics of the true data generating process.

IV. Sample Properties

Thus far we have focused on the properties of the population moments defining the Oaxaca-Blinder estimator. It turns out that the sample moments have some interesting properties as well. Define $N_1=\sum_i D_i$ and $\boldsymbol{X}=\left[\boldsymbol{1},\boldsymbol{x_2},...,\boldsymbol{x_K}\right]$, where $\boldsymbol{1}$ is an $N\times1$ vector of ones and the elements of $\left\{\boldsymbol{x_2},...,\boldsymbol{x_K}\right\}$ are $N\times1$ covariate vectors. Then we may write

[7] This is to be contrasted with the standard logistic assignment model which assumes the odds of treatment take the form $\exp\left(X_i'\gamma\right)$ for some coefficient vector γ. The log-logistic distribution is similar to a log-normal but with heavier tails (the mean of the distribution does not exist). The fact that the support of the distribution is nonnegative is not restrictive as X_i will usually include an intercept.

[8] Both sets of weights can be shown to have mean one which implies $E\left[w\left(X_i\right)-\widetilde{w}\left(X_i\right)|D_i=0\right]=0$.

[9] See Robins, Sued, Lei-Gomez, and Rotnitzky (2007) and Chen, Hong, and Tarozzi (2008) for further discussion of this issue.

the O-B estimate of the counterfactual mean in matrix notation as:

$$\hat{\mu}_0^1 \equiv \frac{1}{N_1}\boldsymbol{D}'\boldsymbol{H}\boldsymbol{Y}$$
$$\boldsymbol{H} \equiv \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{S}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{S}$$

where $\boldsymbol{Y}$ is the $N \times 1$ vector of outcomes, $\boldsymbol{D}$ is an $N \times 1$ vector whose elements consist of D_i, and $\boldsymbol{S}$ is an $N \times N$ diagonal selector matrix taking values equal to $1 - D_i$ along the diagonal and zero elsewhere. The $N \times N$ matrix $\boldsymbol{H}$ is a generalization of the conventional "hat" matrix associated with OLS (Hoaglin and Welch, 1978). Averaging the rows of the hat matrix over the treated observations yields the $1 \times N$ vector of O-B sample weights $\boldsymbol{\omega} \equiv \frac{1}{N_1}\boldsymbol{D}'\boldsymbol{H}$ used to form an estimate $\widehat{\mu}_0^1$ of the average counterfactual outcome in the treated sample. A few properties of these weights are notable:

1) The weights are zero for treated observations.

2) The weights sum to one.[10]

3) Some of the weights may be negative. This occurs when the treatment odds implied by the linear model are negative.

Like conventional propensity score weights, O-B weights can be thought of as reweighting the controls to match the covariate distribution of the treated units. Note that for any covariate $\boldsymbol{x_j}$ in $\boldsymbol{X}$ we have by the properties of projection matrices that:

$$\frac{1}{N_1}\boldsymbol{D}'\boldsymbol{H}\boldsymbol{x_j} = \frac{1}{N_1}\boldsymbol{D}'\boldsymbol{x_j}$$

In words, the reweighted mean of every control covariate exactly equals its mean value among the treated sample. Hence the weights embodied in the Oaxaca-Blinder approach ensure exact balance of moments included in the regression model, a property shared by the recently proposed doubly robust estimator of Egel, Graham, and Pinto (2009).

V. Application

To illustrate use of the Oaxaca-Blinder estimator we revisit LaLonde's (1986) classic analysis of the National Supported Work (NSW) program using observational controls from the Current Population Survey. Attention is confined to a sample of men studied by Dehejia and Wahba (1999) with valid earnings data in both 1974 and 1975 who were present either in the NSW experimental sample or in Lalonde's "CPS-3" control group which consists of the poor and recently unemployed.[11] Because these data have been studied many times, I omit summary statistics which are reported elsewhere.[12] Three estimators: OLS, O-B, and reweighting based upon a logistic propensity score are contrasted; each using the set of demographic controls considered in Dehejia and Wahba (1999) along with 1974 and 1975 earnings.

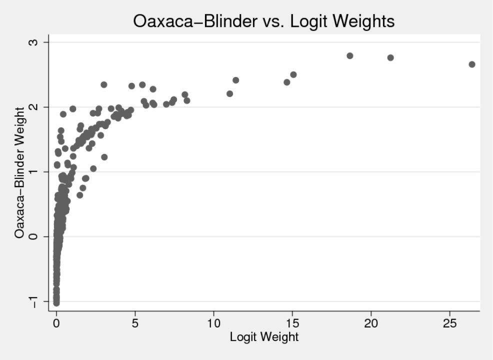

Figure 1

Figure 1 plots a scatter of the renormalized O-B weights (the elements of $\boldsymbol{D}'\boldsymbol{H}$) against the weights $\frac{\widehat{e}(X_i)}{1-\widehat{e}(X_i)}\frac{1-\widehat{\pi}}{\widehat{\pi}}$ derived from a propensity score reweighting estimator where $\widehat{e}(X_i)$ are predicted probabilities from a logit model estimated by Maximum Likelihood and $\widehat{\pi}$ is chosen to ensure the weights sum to N_1 among the controls. Unsurprisingly, the relationship between

[10] Though seemingly mundane, this property may be important in practice. See for example Busso, Dinardo, and McCrary (2009).

[11] See Smith and Todd (2005) for a detailed discussion of the implications of these sample restrictions.

[12] See for example Dehejia and Wahba (1999), Smith and Todd (2005), and Angrist and Pischke (2009).

the two sets of weights is approximately logarithmic. However the O-B weights are often negative, a sign the implicit log-logistic propensity score model is likely misspecified. Of course the logistic model, despite yielding predictions in the unit interval, may also be misspecified. Ultimately, interest centers not on whether a propensity score model is literally correct, but the quality of approximation that can be provided to the true counterfactual μ_0^1.

Table 1 assesses this question empirically by comparing treatment effect estimates generated by each estimator using the observational CPS-3 controls and the experimental NSW controls.[13]

Table 1 - Estimated Impact of NSW on Men's 1978 Earnings		
Estimator/Control Group	CPS-3	NSW
Raw Difference	−$635	$1794
	(677)	(671)
OLS	$1369	$1676
	(739)	(677)
Logistic Reweighting*	$1440	$1808
	(863)	(705)
Oaxaca-Blinder	$1701	$1785
	(841)	(677)
Sample Size	614	445
Note: Heteroscedasticity robust standard errors in parentheses.		
*Reweighting standard errors computed from 1,000 bootstrap replications.		

Clearly covariate adjustments of virtually any sort help to remove bias in the observational sample. However, the O-B estimator yields observational impacts closest to those found in the experimental sample, suggesting the assumption of near linearity of untreated earnings in covariates provides no worse an approximation to the data generating process than the implicit assumptions of the workhorse logistic reweighting estimator. Also of note is that the O-B estimator yields slightly smaller standard error estimates than logistic reweighting, even in the experimental sample.

VI. Conclusion

The regression based Oaxaca-Blinder estimator of counterfactual means is equivalent to a propensity score reweighting estimator modeling the odds of treatment as a linear function of the covariates. This is be to contrasted with the standard practice in the applied literature of modeling the propensity score via a logit or probit and using the estimated parameters to form estimates of the odds of treatment. The latter approach can be thought of as indirectly approximating the unknown odds via a different set of basis functions, albeit a set that imposes the side constraint that the odds are nonnegative. Whether, in the presence of misspecification, the imposition of this side constraint yields a better approximation to the counterfactual of interest is an empirical question and will depend on the data generating process.

Despite its allowance of negative weights, the Oaxaca-Blinder estimator has several features to commend it. It is easily implemented in unbalanced designs with few treated units and many controls and allows for straightforward computation of standard errors and regression diagnostics. It is consistent if either the linear model for the potential outcomes or the implicit log-logistic model for the propensity score is correct. And unlike standard reweighting estimators, the O-B weights yield exact covariate balance and are finite sample unbiased for the counterfactual under proper specification of the outcome equation.

REFERENCES

Angrist, Joshua and Steven Pischke. 2009. *Mostly Harmless Econometrics.* Princeton: Princeton University Press.

Blinder, Alan. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *Journal of Human Resources* 8(4): 436-455.

Busso, Matias, John Dinardo, and Justin McCrary. 2009. "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects." Mimeo.

Busso, Matias, Jesse Gregory, and Patrick Kline. 2010. "Assessing the Incidence and Efficiency of a Prominent Place Based Policy." *NBER Working Paper #16096.*

[13] The Oaxaca-Blinder treatment effect estimator simply subtracts $\hat{\mu}_0^1$ from the mean sample outcome of treated units.

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi. 2008. "Semiparametric Efficiency in GMM models of Nonclassical Measurement Errors, Missing Data, and Treatment Effects." Mimeo.

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

Dinardo, John. 2002. “Propensity Score Reweighting and Changes in Wage Distributions." Mimeo.

DiNardo, John, Nicole Fortin, and Thomas Lemieux. 1996. “Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." *Econometrica* 64(5):1001-1044.

Egel, Daniel, Bryan Graham, and Cristine Pinto. 2009. "Efficient Estimation of Data Combination Problems by the Method of Auxiliary-to-Study Tilting." Mimeo.

Feller, William. 1966. *An Introduction to Probability Theory and Its Applications*. Volume II New York: John Wiley & Sons.

Heckman, James and Richard Robb. 1984. "Alternative Methods for Evaluating the Impact of Interventions." in J. Heckman and B. Singer (eds.) *Longitudinal Analysis of Labor Market Data* Cambridge, U.K.: Cambridge University Press.

Hoaglin, David and Roy Welsch. 1978. "The Hat Matrix in Regression and ANOVA." *The American Statistician* 32(1): 17-22.

Huber, Martin, Michael Lechner, and Conny Wunsch. 2010. "How to Control for Many Covariates? Reliable Estimators Based on the Propensity Score." Mimeo.

Imbens, Guido. 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *Review of Economics and Statistics* 86(1):4-29.

Hirano, Keisuke, Guido Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71(4): 1161-1189.

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4):604-620.

Kang, Joseph and Joseph Schafer. 2007. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical Science* 22(4): 523-539.

Oaxaca, Ronald. 1973. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review* 14(3): 693-709.

Robins, James, Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89(427): 846-866.

Robins, James, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. 2007. "Comment: Performance of Double Robust Estimators When Inverse Probability Weights Are Highly Variable." *Statistical Science* 22(4): 544-559.

Rosenbaum, Paul and Donald Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41-45.

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