Proxy Variable Estimators of Principal Effects

Edward Bein
Abt Associates

October 15, 2013

Abstract

This paper reframes Peck's (2003) approach to estimating impacts on endogenous subgroups in terms of Frangakis and Rubins' (2002) notions of principal stratification and principal effects, and shows her estimators can be derived using proxy variables for the (partly) omitted indicator of principal stratum membership. It presents a specification test and describes a method for estimating bounds for principal effects.

Introduction

Frangakis and Rubin (FR; 2002) introduced the influential notions of principal stratification and principal effects to methodological research on causal inference. However, to paraphrase Moliere, other researchers had been "speaking principal stratification without knowing it" prior to or concurrent with FR's publication. Angrist, Imbens, and Rubin (AIR; 1996) is perhaps the best-known example of this, and their work was discussed by FR. This article focuses on another early "speaker of principal stratification," Peck's (1999, 2003) presentation of two estimators of what are now, via FR, often referred to as principal effects (Peck (2003) did, in fact, discuss FR's work, without making the explicit link). Her estimators were applied to the random assignment evaluation of a welfare reform initiative in New York State, and her methods have subsequently been applied to the evaluation of other social programs (Peck, 2013).

The present article builds upon Peck (2003; hereafter, Peck). First, it reframes Peck's estimands in terms of the language of principal stratification and principal effects. Second, it derives several proxy variable estimators of principal effects, including Peck's two estimators. Peck discussed estimation in terms of instrumental variables, but we believe it is more useful to frame the derivation of estimators in terms of using proxy variables for the (partly) omitted indicator of principal strata membership. Third, it discusses methods for testing and assessing the assumptions underlying the proxy variable estimators, including presenting a specification test similar to tests proposed by Schochet and Burghardt (2007). It also presents the asymptotic bias of one of the proxy estimators, permitting assessment of the sensitivity of the estimates to violations of the validating assumptions. Fourth, it introduces an assumption-free method for estimating upper and lower bounds for principal effects. It concludes with a brief discussion of Bayesian approaches to estimating principal effects.

Peck's Estimands in the Language of Principal Stratification

We present the estimands in terms of the language of principal stratification, adapting notation from FR. For expository purposes, we use the running example of examining, in the context of a random assignment evaluation, the impact of treatment utilization.

Observed Data

Assume a random sample of size *n* from the population of interest, selected to participate in a random assignment evaluation. Let Z_i be the random assignment indicator, where $Z_i = 1$ means participant i was assigned to the treatment group and $Z_i = 0$ means participant i was assigned to the control group. Intermediate variable S_i is the treatment utilization indicator, where $S_i = 1$ means participant i chose to utilize the treatment and $S_i = 0$ means participant i chose not to (in a context different from the running example, S_i might instead be the indicator of which of two versions of the treatment is offered to the participant by program staff). In many evaluations the treatment is not available to control participants and, in such circumstances, $S_i = 0$ for all control participants. Y_i denotes participant i's outcome, and may be continuous or discrete, and X_i is a vector of baseline covariates. It is assumed that Z_i, S_i, Y_i, and X_i are observed for all participants.

Potential Outcomes

Let $S_i(1)$ denote the treatment utilization status that would obtain if, possibly contrary to fact, participant i were randomly assigned to the treatment group. $S_i(1) = 1$ means that participant i would use the treatment, and $S_i(1) = 0$ means that participant i would not use the treatment. $S_i(0)$ is analogous to $S_i(1)$, for the hypothetical situation where participant i is randomly assigned to the control group: $S_i(0) = 1$ means that participant i would use the treatment, and $S_i(0) = 0$ means that participant i would not use the treatment. As noted above, in many randomized experimental evaluations, $S_i(0)=0$ for all participants. $S(1)$ and $S(0)$ are conceptualized as intermediate potential outcomes.

Let $Y_i(1)$ be the outcome that participant i would obtain if, possibly contrary to fact, she were randomly assigned to the treatment group and $Y_i(0)$ is the analogous potential outcome for hypothetical assignment to the control group.

Note the links between observed and potential outcomes: $S_i = S_i(Z_i)$ and $Y_i = Y_i(Z_i)$. That is, if participant i is assigned to the treatment group, then potential outcomes $S_i(1)$ and $Y_i(1)$ are observed but $S_i(0)$ and $Y_i(0)$ go unobserved, and the reverse occurs if the participant is assigned to the control group. Also note that all potential outcomes are (partially observed) baseline covariates, as they are pretreatment characteristics of participants. Because we are focusing on random assignment evaluations, we assume Z_i is independent of the set of baseline covariates (X_i, $S_i(1)$, $S_i(0)$, $Y_i(1)$, $Y_i(0)$).

Principal Strata and Principal Effects

The individual treatment effect is $\Delta_i \equiv Y_i(1) - Y_i(0)$, and represents the impact of being assigned to the treatment group vs. being assigned to the control group for participant i. The average treatment effect (ATE) is a commonly studied causal parameter, and is defined as the population average individual treatment effect

$$ATE \equiv E[\Delta_i] = E[Y_i(1) - Y_i(0)]$$

In the context of a random assignment evaluation, the ATE is the intention-to-treat (ITT) impact of treatment. Because some participants randomized to the treatment group will choose not to actually utilize the treatment (and so will have $S_i = S_i(1) = 0$), and, depending on the design of the evaluation, some randomized to the control group may choose to utilize the treatment (and so will have $S_i = S_i(0) = 1$), the ITT impact does not directly reflect the average impact of treatment utilization. Peck's interest, in the context of the running example, was alternatively in the effect of treatment utilization on those randomized to the treatment group, and can be framed in terms of FR's notions of principal stratification and principal effects, the topic to which I turn now.

Members of the population of interest can be partitioned into four subpopulations based on their $S_i(1)$ and $S_i(0)$ values, where each subpopulation is denoted as *basic principal stratum* BPS(a, b), for a = 0, 1 and b = 0, 1,

$$BPS(a,b) \equiv \{i \mid S_i(1) = a, S_i(0) = b\}$$

where here i indexes members of the population of interest. For example, BPS(0,0) is the subpopulation of individuals who would not utilize the treatment if assigned to the treatment group and also would not utilize it if assigned to the control group. FR termed this 4-way partition the *basic principal stratification* for the population (with respect to S(1) and S(0)). It is possible to create less differentiated partitions of the population, called *principal stratifications*, by combining strata from the basic principal stratification. Peck was interested in the principal stratification, for a = 0, 1,

$$PS(a) \equiv BPS(a,0) \cup BPS(a,1)$$

which combines the four basic principal strata into two strata. Here, PS(1) is the subpopulation of individuals who, if randomized to the treatment group, would utilize the treatment, and PS(0) is the subpopulation of individuals who, if randomized to the treatment group, would not utilize the treatment. In the common but not universal circumstance that it is impossible for control participants to utilize the treatment, BPS(0, 1) and BPS(1, 1) are empty and, for a = 0, 1, PS(a) = BPS(a, 0).

Peck's aim was to estimate and make inferences about the conditional average treatment effects

$$PE(1) \equiv E[\Delta_i \mid i \in PS(1)] = E[\Delta_i \mid S_i(1) = 1]$$

and

$$PE(0) \equiv E[\Delta_i \mid i \in PS(0)] = E[\Delta_i \mid S_i(1) = 0]$$

FR termed causal parameters such as these *principal effects* with respect to principal stratification {PS(1), PS(0)}, because they are average treatment effects defined within principal strata. An analogous pair of principal strata and principal effects could be defined in terms of S(0) rather than S(1), and the same statistical methods presented below would, suitably adjusted, apply to these as well. PE(1) and PE(0) are typically of most substantive interest when it is impossible for control group participants to utilize treatment (i.e., when S(0) is always equal to zero), and Peck assumed this to hold. When this holds, PE(1) is the impact of utilizing vs. not utilizing treatment for the subpopulation that would utilize treatment if they had the chance, and PE(0) is the direct effect of treatment assignment on never-users. In the parlance of AIR, PS(1) is the subpopulation of compliers, and PE(1) is the local average treatment effect (LATE); AIR assumes PE(0) = 0. Note, though, that the principal effect estimators presented below do not rely on the assumption that $S_i(0)=0$ for all i.

Continuous Proxy Variable Estimators

We present three continuous proxy variable estimators of principal effects, dropping the i subscript to limit clutter.

First, let $Prox = Prox(X) \equiv E[S(1)|X]$; Hill, Brooks-Gunn, and Waldfogel (2003) refer to Prox as the *principal score*. Then

$$E[Y|S(1), Z, X, Prox] = E[Y|S(1), Z, X] \quad (1)$$

since Prox is a function of X. Further,

$$E[S(1)|Z, X, Prox] = E[S(1)|X, Prox] = E[S(1)|Prox] = Prox \quad (2)$$

To see the first equality, recall that Z is independent of all combinations of baseline variables. To see the second and third equalities, we adapt an argument from Rosenbaum (2010, p. 73):

$$\begin{aligned} E[S(1)|Prox] &= E\{E[S(1)|X, Prox]|Prox\} = E\{E[S(1)|X]|Prox\} = E[Prox|Prox] = Prox \\ &= E[S(1)|X] = E[S(1)|X, Prox] \end{aligned}$$

If we conceive of S(1) as a (partially) omitted variable, since it is unobserved in the control group, then Prox is a proxy variable for S(1) (Wooldridge, 2010, pp. 67-69):

- (1) indicates that Prox is redundant given omitted variable S(1) (and X and Z) in predicting Y, and
- (2) indicates that Z and X are redundant given proxy Prox in determining the linear projection of S(1) on Prox, Z, and X.

Assume

$$E[Y|S(1),Z,X] = \beta_0 + \beta_1 S(1) + \beta_2 Z + \beta_3 S(1)Z + \beta_4 X + \beta_5 ZX + \beta_6 S(1)X + \beta_7 S(1)ZX \quad (3)$$

Because Z is independent of all baseline variables (including S(1)),

$E[Y|S(1), Z = 1, X] = E[Y(1)|S(1), Z = 1, X] = E[Y(1)|S(1), X]$,

$E[Y|S(1), Z = 0, X] = E[Y(0)|S(1), Z = 0, X] = E[Y(0)|S(1), X]$, and

$$\begin{aligned} E[\Delta|S(1), X] &= E[Y|S(1), Z = 1, X] - E[Y|S(1), Z = 0, X] \\ &= \beta_2 + \beta_3 S(1) + \beta_5 X + \beta_7 S(1)X \end{aligned} \quad (4)$$

Thus,

$$E[\Delta|S(1) = 1, X] = \beta_2 + \beta_3 + (\beta_5 + \beta_7)X$$

and

$$E[\Delta|S(1) = 0, X] = \beta_2 + \beta_5 X \quad (5)$$

Linear regression (3) cannot be fit on the full sample, because S(1) is (partly) omitted. However, applying (1) and (2) to (3), we obtain

$$E[Y|Prox, Z, X] = \beta_0 + \beta_1 Prox + \beta_2 Z + \beta_3 ProxZ + \beta_4 X + \beta_5 ZX + \beta_6 ProxX + \beta_7 ProxZX \quad (6)$$

Via fitting a logistic regression of S(1)=S on X on the treatment subsample, generated regressor $\widehat{Prox}$ can be determined for all sample participants, and then linear regression (6), with $\widehat{Prox}$ in place of Prox, can be fit on the entire sample. Via the (6) regression coefficient estimates, estimates of PE(1) and PE(0) can be computed based on (5) and the fact that, by iterated expectation, for a=0, 1,

$$PE(a) = E\{E[\Delta|S(1) = a, X]\}$$

where the outer expectation is with respect to the conditional distribution of X given S(1)=a; i.e., the distribution of X within PS(a). Letting $\overline{X^{(1)}}$ denote the mean X for the subsample of treatment participants who have S(1)=1 and $\overline{X^{(0)}}$ denote the mean X for the subsample of treatment participants who have S(1)=0, we get

$$\widehat{PE}(1) = \widehat{\beta_2} + \widehat{\beta_3} + (\widehat{\beta_5} + \widehat{\beta_7})\overline{X^{(1)}}$$

and

$$\widehat{PE}(0) = \widehat{\beta_2} + \widehat{\beta_5}\overline{X^{(0)}}$$

However, there are good reasons to doubt the adequacy of using (6) given likely issues with multicollinearity. To illustrate this, imagine that $\widehat{Prox}$ was generated via a linear probability model rather than a logistic regression model. In this circumstance, β_3 and β_5 are not identified, since $\widehat{Prox}$ is a linear combination of the components of X (and 1). Since estimates of these two regression coefficients are needed to estimate both principal effects, the use of (6) would be doomed. Using a nonlinear model like logistic regression to generate $\widehat{Prox}$ for use in an assumed linear model obviates the possibility of perfect collinearity, but problems with multicollinearity would not be unexpected. If this occurred, then extremely large samples might be needed to obtain precise estimates of the principal effects. The use of ridge regression or the lasso (James, Witten, Hastie, & Tibshirani, 2013), rather than OLS, for fitting (6) could then be considered as approaches for dealing with the multicollinearity.

A second approach to using (6) is to assume that some of the regression coefficients in (3) are equal to zero. This, in effect, is the approach taken by Jo and Stuart (2009) with their assumption of principal ignorability. In the present context, consider the following ``abbreviated’’ version of (3)

$$E[Y|S(1), Z, X] = \beta_0 + \beta_1 S(1) + \beta_2 Z + \beta_3 S(1)Z + \beta_4 X + \beta_6 S(1)X \tag{7}$$

This leads to the ``abbreviated’’ version of (6)

$$E[Y|Prox, Z, X] = \beta_0 + \beta_1 Prox + \beta_2 Z + \beta_3 ProxZ + \beta_4 X + \beta_6 ProxX \tag{8}$$

Now we have

$$E[\Delta|S(1), X] = \beta_2 + \beta_3 S(1) \tag{9}$$

There are no multicollinearity concerns with the regressors associated with β_2 or β_3, the only regression coefficients involved in (9). In place of (8), the following model should be fit

$$E[Y|Prox, Z] = \beta_0 + \beta_1 Prox + \beta_2 Z + \beta_3 ProxZ + \beta_4 E[X|Prox] + \beta_6 ProxE[X|Prox]$$

or, better, fit the following to obtain estimates of β_2 and β_3

$$E[Y|Prox, Z] = \beta_0 + \beta_{146} h(Prox) + \beta_2 Z + \beta_3 ProxZ \tag{10}$$

where h(Prox) is some suitable (linear or nonlinear) function of Prox. In (10) we are unable to distinguish the contributions of β_1, β_4, and β_6, but these are not needed to estimate the principal effects.

Equation (9) implies that $E[\Delta|S(1),X] = E[\Delta|S(1)]$; i.e., X is redundant given S(1). Per (7), X may be associated with Y(1) and/or Y(0) given S(1), but it is unassociated with their difference given S(1).

We now give a third continuous proxy variable estimator, originally presented by Peck. Assume

$$E[Y|S(1),Z,Prox] = E[Y|S(1),Z] \tag{11}$$

That is, Prox is assumed redundant in predicting Y given S(1) and Z. Then, given (2), Prox is a continuous proxy variable for S(1). Consider the saturated (and hence correctly specified) linear regression model

$$E[Y|S(1),Z] = \alpha_0 + \alpha_1 S(1) + \alpha_2 Z + \alpha_3 S(1)Z \tag{12}$$

Per this equation,

$$PE(1) = E[Y|S(1) = 1, Z = 1] - E[Y|S(1) = 1, Z = 0] = \alpha_2 + \alpha_3$$
$$PE(0) = E[Y|S(1) = 0, Z = 1] - E[Y|S(1) = 0, Z = 0] = \alpha_2$$

Note, in passing, that AIR do not make redundancy assumption (11) but do assume, in (12), that $\alpha_2 = 0$. This follows from their assumptions that there are no defiers and no direct effect of treatment assignment among never-takers.

By assumption (11), and using (2),

$$E[Y|Prox,Z] = E\{\, E[Y|S(1),Z,Prox]|Prox,Z\} = \alpha_0 + \alpha_1 Prox + \alpha_2 Z + \alpha_3 ProxZ \tag{13}$$

Fitting (13), with generated regressor $\widehat{Prox}$ in place of Prox, yields consistent estimates of the regression coefficients, from which the principal effects estimates can be computed.

In summary, all three continuous proxy variable estimators are derived from fitting an appropriate linear regression using generated regressor $\widehat{Prox}$, and the principal effects estimates are linear combinations of the regression coefficient estimates. Since the regression coefficient estimators are asymptotically multivariate normal, the principal effects estimators are asymptotically normal. As Peck noted, because a generated regressor is used, the estimation proceeds via a 2-step procedure (the logistic regression used to generate $\widehat{Prox}$ is fit at step 1, OLS estimation of regression coefficients is step 2) which yields different standard errors than if the true values of Prox were used. The bootstrap can be used to account for the 2-step procedure in estimating standard errors.

A Binary Proxy Variable Estimator

Let $BProx = BProx(X)$ be some binary variable that is a function of X (to be discussed below). Note

$$E[S(1)|Z, BProx] = E[S(1)|BProx] = \gamma_0 + \gamma_1 \, BProx$$

since Z is independent of all baseline variables and the right-hand side of the second equality is a saturated model and hence is correctly specified. This can equivalently be written

$$S(1) = \gamma_0 + \gamma_1 \, BProx + r \qquad (14)$$

Assume

$$E[Y|S(1), Z, BProx] = E[Y|S(1), Z] \qquad (15)$$

Then BProx is a proxy variable for S(1), since it is redundant in predicting Y and Z is redundant in predicting S(1).

Now reconsider the saturated (and hence correctly specified) linear regression model (12), which can be written in error form as

$$Y = \alpha_0 + \alpha_1 S(1) + \alpha_2 Z + \alpha_3 S(1) Z + e$$

By assumption (15), BProx is uncorrelated with e. Substituting for S(1) in the linear regression, we obtain

$$Y = \pi_0 + \pi_1 BProx + \pi_2 Z + \pi_3 BProxZ + \varepsilon \qquad (16)$$

where

$$\begin{gathered} \pi_0 = \alpha_0 + \alpha_1 \gamma_0 \\ \pi_1 = \alpha_1 \gamma_1 \\ \pi_2 = \alpha_2 + \alpha_3 \gamma_0 \\ \pi_3 = \alpha_3 \gamma_1 \\ \varepsilon = \alpha_1 r + \alpha_3 r Z + e \end{gathered}$$

Solving for the regression coefficients from (12), we get (17)

$$\begin{gathered} \alpha_0 = \pi_0 - \pi_1 \gamma_0 / \gamma_1 \\ \alpha_1 = \pi_1 / \gamma_1 \\ \alpha_2 = \pi_2 - \pi_3 \gamma_0 / \gamma_1 \\ \alpha_3 = \pi_3 / \gamma_1 \end{gathered}$$

We note that BProx, Z, and BProxZ are all exogenous in (16). This suggests the following procedure:

1. Regress S(1)=S on BProx on the treatment group to obtain $\widehat{\gamma_0}$ and $\widehat{\gamma_1}$.
2. Regress Y on BProx, Z, and BProxZ on the entire sample, obtaining estimates of the π coefficients.
3. Estimates of the α coefficients are obtained from (17) and the estimates from the first two steps.
4. The estimates of the primary effects are obtained from the alpha estimates.

Note that we can characterize the principal effects in terms of the πs and γs as:

$$PE(0) = \frac{\gamma_1\pi_2 - \gamma_0\pi_3}{\gamma_1}$$

and

$$PE(1) = PE(0) + \pi_3/\gamma_1$$

As shorthands, let P_{ab} denote $P(S(1) = a|BProx = b)$ and E_b denote $E[\Delta|BProx = b]$. Noting that, for instance, $\gamma_0 = P_{10}$, $\pi_0 = E[Y(0)|BProx = 0]$, and so on, we find that

$$PE(0) = \frac{P_{11}E_0 - P_{10}E_1}{P_{11} - P_{10}}$$

and (18)

$$PE(1) = \frac{P_{00}E_1 - P_{01}E_0}{P_{11} - P_{10}}$$

Replacing the terms in the numerators and denominators by their sample analogues yields the estimators for the principal effects. That is,

$$\hat{P}_{ab} = \hat{P}(S = a|BProx = b, Z = 1)$$

and

$$\hat{E}_b = \hat{E}[Y|BProx = b, Z = 1] - \hat{E}[Y|BProx = b, Z = 0]$$

where the terms on the right-hand sides are subsample means.

These expressions for the principal effects were obtained by Peck, but relying on a weaker assumption than (15). Peck noted that, by iterated expectation, we have

$$E[\Delta|BProx = 1] = P_{11}E[\Delta|S(1) = 1, BProx = 1] + P_{01}E[\Delta|S(1) = 0, BProx = 1]$$

and

$$E[\Delta|BProx = 0] = P_{10}E[\Delta|S(1) = 1, BProx = 0] + P_{00}E[\Delta|S(1) = 0, BProx = 0]$$

Make the redundancy assumption, implied by (15), that

$$E[\Delta|S(1)] = E[\Delta|S(1), BProx] \tag{19}$$

That is, assume that

$$E[\Delta|S(1) = 1, BProx = 0] = E[\Delta|S(1) = 1] = E[\Delta|S(1) = 1, BProx = 1]$$

and that

$$E[\Delta|S(1) = 0, BProx = 0] = E[\Delta|S(1) = 0] = E[\Delta|S(1) = 0, BProx = 1]$$

Under assumption (19) we have

$$E[\Delta|BProx = 1] = P_{11}PE(1) + P_{01}PE(0)$$

and

$$E[\Delta|BProx = 0] = P_{10}PE(1) + P_{00}PE(0)$$

Now solving for PE(1) and PE(0) gives (18).

The binary proxy variable estimators given in (18) are asymptotically normally distributed; see the Appendix.

Finally, suppose that we had two binary proxy variables, BProx1 and BProx2, both satisfying assumption (19). Then consistent principal effects estimators could be derived from either of them. If BProx1 is more strongly correlated with S(1) than is BProx2, then r1 (the error term when (14) is fit using BProx1) will have a smaller variance than will r2 (the error term when (14) is fit using BProx2). This, in turn, implies that ε in (16) will have smaller variance if BProx1 rather than BProx2 is used. That is, (16) will be a better-fitting model if BProx1 is used instead of BProx2. Now, one way to produce a putative binary proxy variable BProx from X is to fit a logistic regression of S(1) on X in the treatment subsample, generate predicted probabilities of S(1)=1 for all participants in the sample, and then set a cutoff probability below which BProx is set to 0 and above which it is set to 1. Using .5 as the cutoff produces an estimated Bayes classifier, which maximizes the proportion of correct classifications (James, Witten, Hastie, & Tibshirani, 2013). That is, the Bayes classifier, defined as applying cutoff .5 to the true probabilities that S(1)=1, maximizes the proportion of the population for which either BProx=1=S(1) or BProx=0=S(1) holds. However, in general, BProx implemented using a cutoff of .5 will not produce a binary variable that is maximally correlated with S(1); often using a different cutoff produces a binary variable that is more strongly correlated with S(1) and hence produces sharper principal effects estimates.

Assessing Assumptions

This section begins by presenting a specification test that can be used to assess the assumptions underlying any of the proxy variable estimators of PE(1) and PE(0) described above; indeed, it can be applied to any estimators of PE(1) and PE(0). We then consider how to assess the redundancy assumptions underlying the two proxy variable estimators presented by Peck: (a) redundancy assumption (11) underlying the continuous proxy variable estimator derived from linear regression model (13) and (b) redundancy assumption (19) underlying the binary proxy variable estimator given in (18).

A Specification Test

It is possible to test the assumptions underlying a given estimator (whether based on proxy variables or not) of the principal effects by estimating the average treatment effect in two different ways, one that is consistent whether or not the underlying assumptions hold and the other that is consistent under the assumptions. First, the ATE can be given by

$$ATE = E[Y|Z = 1] - E[Y|Z = 0]$$

which leads to the "natural" estimator of the ATE by using the appropriate sample means in place of the population means. The natural estimator is consistent and asymptotically normal whether or not the assumptions underlying the (proxy variable) estimator hold.

Second, the ATE can also be characterized

$$ATE = P(S(1) = 1)PE(1) + P(S(1) = 0)PE(0)$$

This leads to a proxy variable estimator of the ATE by using one of the proxy variable estimators of the principal effects in place of the principal effects (note $P(S(1) = 1)$ and $P(S(1) = 0)$ can be estimated from the treatment subsample). The resulting proxy variable estimator is a consistent and asymptotically normal estimator of the ATE under the appropriate assumptions. Let $diff \equiv \widehat{ATE}_{nat} - \widehat{ATE}_{proxy}$ represent the difference between the natural and proxy variable ATE estimators. If the appropriate assumptions hold, then the test statistic

$$\frac{diff}{SE(diff)}$$

is asymptotically standard normal. The bootstrap can be used to estimate SE(diff). Further, a bootstrap-within-the-bootstrap approach can be used to implement a percentile-t testing procedure that achieves asymptotic refinement (Cameron & Trivedi, 2005, pp. 378-379). We reiterate that the bootstrap procedures should include the generation of the proxy variable regressor to account for the two-step nature of the proxy variable estimators.

Note that it is possible that proxy variable estimators for PE(1) and PE(0) are asymptotically biased due to violation of their underlying assumptions and, at the same time, that $\widehat{ATE}_{proxy}$ is consistent due to the biases offsetting one another. Thus, the power of the specification test may be extremely low at some alternatives; indeed, if the biases exactly offset one another, asymptotic power is equal to the size of the test.

Assessing the Redundancy Assumption for the Continuous Proxy Variable Estimator

Redundancy assumption (11), positing that Prox is redundant in predicting Y given S(1) and Z, implies the specific functional form for $E[Y|Prox, Z]$ that is given in (13). If this functional form is misspecified, then (11) is false. Therefore, the usual repertoire of statistical methods for testing or assessing functional form can be employed to indirectly test or assess (11), and these provide an alternative or adjunct to using the specification test. For example, a model comparison test of (13) vs. a model that additionally includes polynomial terms for $Prox$, examination of residual plots for (13), and using cross-validation to compare (13) to alternative regression models with respect to MSE would all be ways to indirectly test or assess (11). A complication here is that $Prox$ itself is not used to fit (13), but rather a version of $Prox$ generated from a model for $E[S(1)|X]$, and an apparent inadequacy of (13) may in fact be due to a misspecified model for $E[S(1)|X]$. In practice, then, care must be taken in specifying and checking the model for $E[S(1)|X]$ before examining possible inadequacies of (13).

Assessing the Redundancy Assumption for the Binary Proxy Variable Estimator

The binary proxy variable principal effects estimator can be derived from fitting linear regression models (14) and (16) or from direct substitution into (18), and the estimator relies on redundancy assumption (19), which posits that BProx is redundant in predicting the individual treatment effect given S(1). However, both (14) and (16) are necessarily correctly specified, whether or not (19) holds, and therefore the kind of model-misspecification-based assumption testing or checking suitable for the continuous proxy variable estimator is unsuitable in the binary proxy variable case. If X, or some subvector of X, is conditionally independent of Δ given S(1), then any binary variable derived from X (or its subvector) satisfies (19): conditional independence of X implies the weaker assumption of the redundancy of X, which implies the redundancy of any binary variable derived from X. In some evaluations, there may be good substantive grounds for arguing that a given subvector of X meets the conditional independence assumption or redundancy assumption, while in other evaluations substantive knowledge may undermine this supposition. In either circumstance, the specification test can be used to assess (19).

The Asymptotic Bias of the Binary Proxy Variable Estimator

As an additional tool for working with the binary proxy variable estimator, we present its asymptotic bias. This bias can be characterized in terms of two parameters, δ_1 and δ_0, that are equal to zero if redundancy assumption (19) holds. Violations of (19) can be specified by ascribing nonzero values to at least one of these parameters, and the magnitude of the consequent bias can be examined to determine the magnitude of the assumption violation needed to change the substantive interpretation of the primary effects estimates. Thus, the asymptotic bias formulas can be used to assess the sensitivity of the principal effects estimates to violations of the underlying redundancy assumption.

As a shorthand, let $E_{ab} = E[\Delta|S(1) = a, BProx = b]$. Then we can write

$$E_{11} = E_{10} + \delta_1$$

and

$$E_{01} = E_{00} + \delta_0$$

If δ_1 and δ_0 are both zero, then (19) holds, and otherwise (19) is false; the sizes of δ_1 and δ_0 indicate the magnitude of the divergence from (19). From Peck's derivation of the binary proxy variable estimator, we have

$$E_1 = P_{11}(E_{10} + \delta_1) + P_{01}(E_{00} + \delta_0)$$

$$E_0 = P_{10}E_{10} + P_{00}E_{00}$$

allowing us to solve for E_{10} and E_{00}. Additionally, by iterated expectation we obtain

$$PE(0) = E_{00} + P(BProx = 1|S(1) = 0)\delta_0$$

$$PE(1) = E_{10} + P(BProx = 1|S(1) = 1)\delta_1$$

The binary proxy variable estimators consistently estimate, respectively,

$$QE(0) \equiv \frac{P_{11}E_0 - P_{10}E_1}{P_{11} - P_{10}}$$

and

$$QE(1) \equiv \frac{P_{00}E_1 - P_{01}E_0}{P_{11} - P_{10}}$$

where, under (19), QE(0) = PE(0) and QE(1)=PE(1) but not otherwise. Then the asymptotic bias of $\widehat{PE}(0)$ is

$$PE(0) - QE(0) = \frac{P_{10}P_{11}\delta_1 + P_{10}P_{01}\delta_0}{P_{11} - P_{10}} + P(BProx = 1|S(1) = 0)\delta_0$$

and the asymptotic bias of $\widehat{PE}(1)$ is

$$PE(1) - QE(1) = -\frac{P_{00}P_{11}\delta_1 + P_{00}P_{01}\delta_0}{P_{11} - P_{10}} + P(BProx = 1|S(1) = 1)\delta_1$$

Since the P_{ab} and $P(BProx = 1|S(1))$ can be estimated from the treatment subsample, for any specified values of δ_1 and δ_0 the asymptotic biases of the principal effects estimators can be estimated.

Estimating Bounds for the Principal Effects

This section presents an assumption-free method for estimating upper and lower bounds for principal effects PE(1) and PE(0). Note that

$$E[\Delta|S(1)] = \beta_0 + \beta_1 S(1)$$

is a saturated, and hence correctly specified, linear model. Using a standard linear regression result (e.g., Wooldridge, 2010, p. 25),

$$\beta_1 = \frac{Cov(\Delta, S(1))}{Var(S(1))}$$

and (20)

$$\beta_0 = E[\Delta] - E[S(1)]\beta_1 = ATE - E[S(1)]\beta_1$$

Here, $PE(0) = \beta_0$ and $PE(1) = \beta_0 + \beta_1$. All of the terms on the right-hand sides can be estimated from sample data except for $Cov(\Delta, S(1))$. Hence, finding bounds for $Cov(\Delta, S(1))$ gives us bounds for the principal effects. We now describe an approach to finding bounds for $Cov(\Delta, S(1))$.

To start,

$$Cov\big(\Delta, S(1)\big) = Cov\big(Y(1), S(1)\big) - Cov(Y(0), S(1))$$

where the first term on the right-hand side is equal to $Cov(Y, S|Z = 1)$ and hence can be estimated. For the second term on the right-hand side, we have the inequality

$$-SD\big(Y(0)\big)SD(S(1)) \leq Cov(Y(0), S(1)) \leq SD\big(Y(0)\big)SD(S(1))$$

implying that

$$\begin{aligned} Cov(Y,S|Z=1) - SD(Y|Z=0)SD(S|Z=1) &\leq Cov\big(\Delta, S(1)\big) \\ &\leq Cov(Y,S|Z=1) + SD(Y|Z=0)SD(S|Z=1) \end{aligned}$$

The upper and lower bounds can thus be estimated from the sample data. These estimated bounds can be applied to (20) to obtain estimated upper and lower bounds for the principal effects.

Discussion

AIR and Peck, as well as principal effect estimation utilizing weights derived from principal scores (e.g., Jo & Stuart, 2009), are frequentist methods that rely on differing identifying assumptions, and the specification test presented above can be employed to test those assumptions. We conclude by noting that Bayesian methods can be employed when the principal effects are not identified by the observed data. Rubin (2005, pp. 326-327) provides a concise description of Bayesian methods for causal inference, and Taylor and Zhou (2009) is an example of using multiple imputation for inference about principal effects. A useful next research step involves investigating the relative merits of Bayesian inference for principal effects vis-a-vis the frequentist methods of Peck and others.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444-455.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. New York: Cambridge University Press.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*, 21-29.

Hill, J. L., Brooks-Gunn, J., & Waldfogel, J. B. B. (2003). Sustained effects of high participation in an early intervention for low-birthweight premature infants. *Developmental Psychology*, *39*, 730-744.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York: Springer.

Jo, B., & Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*, *28*, 2857-2875.

Peck, L. R. (1999). *Do social programs affect nonparticipants? Evidence from the child assistance program*. Paper presented at the New York University Robert F. Wagner Graduate School of Public Service Doctoral Colloquium, New York, NY.

Peck, L. R. (2003). Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice. *American Journal of Evaluation*, *24*, 157-187.

Peck, L. R. (2013). On analysis of symmetrically predicted endogenous subgroups: Part one of a method note in three parts. *American Journal of Evaluation*, *34*, 225-236.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322-331.

Schochet, P. Z., & Burghardt, J. (2007). Using propensity scoring to estimate program-related subgroup impacts in experimental program evaluations. *Evaluation Review*, *31*, 95-120.

Taylor, L., & Zhou, X. H. (2009). Multiple imputation methods for treatment noncompliance and nonresponse in randomized clinical trials. *Biometrics*, *65*, 88-95.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data (2nd ed.)*. Cambridge, MA: MIT Press.

Appendix

We present the large-sample distribution of the binary proxy variable estimator of PE(1), denoted $\widehat{PE}(1)$, that is given in (18). Let

$$w_1 = \frac{P_{00}}{P_{11} - P_{10}}$$

$$w_0 = \frac{P_{01}}{P_{11} - P_{10}}$$

$$I_{zb} = I(Z = z, BProx = b)$$

$$p_{zb} = P(I_{zb} = 1)$$

$$M_{zb} = E[Y|I_{zb} = 1]$$

$$V_1 = \frac{M_{11}^2(1 - p_{11}) + Var(Y|I_{11} = 1)}{p_{11}} + \frac{M_{01}^2(1 - p_{01}) + Var(Y|I_{01} = 1)}{p_{01}} + 2M_{11}M_{01}$$

$$V_0 = \frac{M_{10}^2(1 - p_{10}) + Var(Y|I_{10} = 1)}{p_{10}} + \frac{M_{00}^2(1 - p_{00}) + Var(Y|I_{00} = 1)}{p_{00}} + 2M_{10}M_{00}$$

Then it can be shown that

$$\sqrt{n}(\widehat{PE}(1) - PE(1)) \overset{d}{\rightarrow} N(0, w_1^2 V_1 + w_0^2 V_0 + 2w_1 w_0 E_1 E_0)$$

Hence, in sufficiently large samples, $\widehat{PE}(1)$ is approximately normally distributed with mean PE(1) and variance

$$\frac{w_1^2 V_1 + w_0^2 V_0 + 2w_1 w_0 E_1 E_0}{n}$$

This corrects the estimator variance reported in Peck (2003, p. 184). Analogous results hold for the binary proxy variable estimator of PE(0). An important caveat: these results do not account for the fact that *BProx* will typically be a generated regressor. Therefore, using the bootstrap to get standard errors will likely be more accurate than relying on the results above.