Harvard University

Harvard University Biostatistics Working Paper Series

Year 2014 *Paper* 175

A Note on the Control Function Approach with an Instrumental Variable and a Binary Outcome

Eric Tchetgen Tchetgen[*]

[*]Harvard School of Public Health, etchetge@hsph.harvard.edu

Commentary

A note on the control function approach with an instrumental variable and a binary outcome

Eric J. Tchetgen Tchetgen[1,2]

Departments of Biostatistics and Epidemiology,

Harvard School of Public Health

Corresponding author: Eric J. Tchetgen Tchetgen, Department of Epidemiology, Harvard School of Public Health 677 Huntington Avenue, Boston, MA 02115.

Running head:control function for binary exposure

KEY WORDS: Unobserved confounding, instrumental variable, control function, logistic regression

Word count, main text:1796

Word count, abstract:248

Number of Figures:0

Number of tables:0

Number of Pages:14

Abstract

Unobserved confounding is a well known threat to causal inference in non-experimental studies. The instrumental variable design can under certain conditions be used to recover an unbiased estimator of a treatment effect even if unobserved confounding cannot be ruled out with certainty. For continuous outcomes, two stage least squares is the most common instrumental variable estimator used in epidemiologic applications. For a rare binary outcome, an analogous linear-logistic two-stage procedure can be used. Alternatively, a control function approach is sometimes used which entails entering the residual from the first stage linear model as a covariate in a second stage logistic regression of the outcome on the treatment. Both strategies for binary response have previously formally been justified only for continuous exposure, which has impeded widespread use of the approach outside of this setting. In this note, we consider the important setting of binary exposure in the context of a binary outcome. We provide an alternative motivation for the control function approach which is appropriate for binary exposure, thus establishing simple conditions under which the approach may be used for instrumental variable estimation when the outcome is rare. In the proposed approach, the first stage regression involves a logistic model of the exposure conditional on the instrumental variable, and the second stage regression is a logistic regression of the outcome on the exposure adjusting for the first stage residual. In the event of a non-rare outcome, we recommend replacing the second stage logistic model with a risk ratio regression.

In recent years, the instrumental variable (IV) design has gained popularity in epidemiology, as a strategy to recover unbiased estimates of an exposure or treatment causal effect in settings where unobserved confounding is suspected to be present (Greenland, 2000, Davey Smith and Ebrahim, 2003, Hernán and Robins, 2006, Lawlor et al, 2008, Palmer et al, 2011). For continuous outcomes, the most common IV estimator used in practice is two stage least squares which involves fitting a linear regression of the outcome on an estimate of the exposure mean, obtained by regressing the exposure on the IV in a first stage linear model (Wooldridge, 2002). For a rare binary outcome an analogous two stage procedure is sometimes used, in which linear regression is used in the first stage, however, a logistic regression is substituted in the second stage to account for the binary nature of the outcome (Theil, 1953, Basmann, 1957, Angrist, 2001, Wooldridge, 2002, Didelez et al, 2010). A variation of the approach simply adjusts for the residual of the first stage linear regression of the exposure, in a second stage logistic regression of the outcome on the observed treatment; this strategy is described in the literature as a control function approach (Garen, 1994, Woldridge, 1997, Nagelkerke et al, 2000, Blundell and Powell, 2003, Terza et al, 2008). Both strategies for binary response have previously formally been justified only for continuous exposure (Mullahy,1997, Didelez, 2010, Vansteelandt et al, 2011), which has impeded widespread application of either method outside of this setting. In this note, we consider the important setting of binary treatment in the context of a binary outcome. We provide an alternative formulation of the second strategy which applies for binary exposure, thus establishing simple conditions under which the approach may be used for instrumental variable estimation when the outcome is rare and the treatment is binary. In the proposed control function approach, the first stage regression involves a logistic model of the exposure conditional on the instrumental variable, and the second stage regression is a logistic regression of the outcome on the exposure adjusting for the first stage residual. In the event of a non-rare outcome, we recommend replacing the second stage with a risk ratio regression.

Review of control function for continuous treatment

Suppose that one has observed a rare binary outcome Y, a continuous exposure A, and a binary instrumental variable Z. Throughout, U will refer to an unmeasured continuous confounder of the A-Y causal association. A standard formulation of a data generating model for the control function approach assumes the outcome is generated from the log linear model

$$\log \Pr\left(Y=1|A,Z,U\right)=\beta_0+\beta_A A+U, \tag{1}$$

with β_A the log risk ratio causal association between A and Y conditional on U (see for example Palmer et al, 2011). The model rules out the possibility of latent effect heterogeneity of A wrt U on the multiplicative scale. For continuous A, the model further posits the following model for the exposure :

$$A=\gamma_0+\gamma_1 Z+\gamma_2 U+\varepsilon \text{ where } \varepsilon \text{ is independent mean zero error,} \tag{2}$$

where $\gamma_2 \neq 0$. In addition the model assumes that U and Z are independent, as would be the case for a valid IV. Note that the above model encodes explicitly the assumption that Z is a valid instrumental variable, which satisfies the following conditions:

1. Z only affect Y through its association with A, which is encoded by the fact that although Z appears in the conditioning event on the left hand side of equation (1), it does not appear on the right hand side of the equation.

2. The unmeasured confounder of the exposure effect on the outcome is independent of the IV, thus Z is independent of U .

3. The IV is relevant for the exposure, i.e. Z predicts A and thus $\gamma_1 \neq 0$.

Note that this formulation assumes U does not interact with Z in the model for A. Under these assumptions, one can show that

$$\log \Pr(Y = 1|A, Z) = \beta_0^* + \beta_A A + \alpha_1 \Delta$$

where

$$\Delta = A - E\left(A|Z\right).$$

The above equation gives a simple parametrization for the log-linear regression of Y on (A, Z), which allows one to recover under the stated assumptions, the causal log risk ratio association β_A between A and Y. Estimation typically proceeds in two stages. In the first stage, one fits a standard linear regression to estimate the exposure model

$$E\left(A|Z\right) = \alpha_0 + \alpha_1 Z$$

by ordinary least squares (ols), which in turn is used to estimate the residual Δ with $\widehat{\Delta} = A - \left(\widehat{\alpha}_0 + \widehat{\alpha}_1 Z\right)$, where $(\widehat{\alpha}_0, \widehat{\alpha}_1)$ are ols estimates. Then in a second stage, one regresses Y on $(A, \widehat{\Delta})$ using standard logistic regression, as a suitable approximation for the log-linear model (4). The regression coefficient for the exposure in the second stage logistic regression will then be approximately unbiased for β_A. We will refer to the above two stage procedure as the linear-logistic control function approach. The large sample variance of the resulting estimator of β_A must acknowledge the first stage estimation of $E\left(A|Z\right)$, which is easily obtained from standard M-estimation theory. Alternatively, when convenient, one could also use the nonparametric bootstrap to obtain confidence intervals.

One can assess the extent of unobserved confounding, by evaluating the strength of association between Y and Δ, which can be performed with a test of the null hypothesis that $\alpha_1 = 0$.

Control function for binary treatment

Now, suppose that A is dichotomous, then as noted by Didelez et al (2010), assumption (2) cannot be satisfied for binary exposure. Thus, we will consider an alternative formulation, whereby assumption (2) is replaced by the following location shift model for U :

$$U = E\left(U|A, Z\right) + \delta \text{ where } \delta \text{ is independent of } (A, Z) \tag{3}$$

The assumption would hold if U were normally distributed given (A, Z) with homoscedastic variance, however, this is not strictly required and the model allows for an arbitrary distribution for δ.

Result 1: Under assumptions (1),(3) , *and the assumption that U and Z are independent, we have that,*

$$\log \Pr(Y = 1|A, Z) = \beta_0^* + \beta_A A + \ \left(\omega_1 + \omega_2 Z\right) \Delta, \tag{4}$$

where $(\beta_0^*, \omega_1, \omega_2)$ are defined in the appendix.

Result 1 provides formal justification for a generalization of the control function approach in the context of a binary treatment, with the standard control function approach recovered by setting $\omega_2 = 0$. Interestingly, unlike the standard formulation for continuous treatment, the regression model (4) formally allows for heterogeneity in the degree of selection bias due to confounding if $\omega_2 \neq 0$.

Implementation of the approach in practice is fairly straightforward. The main adjustment to account for binary treatment is in the first stage estimation of the treatment model, whereby ols estimation of a linear model for continuous treatment can be replaced with maximum likelihood estimation (mle) of a logistic regression for binary treatment:

$$\text{logit}\Pr(A=1|Z)=\alpha_0+\alpha_1 Z$$

to produce an estimated propensity score $\widehat{\pi}(Z)=\widehat{\Pr}(A=1|Z)$ using the mle $(\widehat{\alpha}_0,\widehat{\alpha}_1)$. The second stage of the approach proceeds by estimating the logistic regression of Y on $(A,\widehat{\Delta},Z\widehat{\Delta})$, upon redefining the estimated residual as $\widehat{\Delta}=A-\widehat{\pi}(Z)$. The resulting estimator of the regression coefficient for the exposure in the second stage logistic regression will be approximately unbiased for β_A provided that the assumptions of Result 1 hold. For inference, one may use M-estimation theory to derive the large sample variance of the estimator, alternatively, one can proceed with the nonparametric bootstrap.

Control function when the outcome is not rare

If Y is not rare in the target population, one may adopt one of several existing methods to estimate the risk ratio regression (4), including the log-binomial model of Wacholder (1986), the Poisson regression approach of Zou (2004), and the semiparametric locally efficient approach of Tchetgen Tchetgen (2013).

Control function under case-control sampling

Case-control studies are a common design in epidemiologic practice, particularly is settings where Y is rare in the population, or measuring Z or A is costly. Accounting for case-control ascertainment is fairly straightforward in the case of a rare outcome, since logistic regression, which appropriately accounts for the sampling design can continue to be used in the second stage, however, the first stage regression model must be modified to account for possible selection bias. A simple strategy entails restricting estimation of the first stage regression of A on Z, to the subset of controls with

$Y = 0$, which should yield a reasonable approximation of the population regression model. This approach may however be inefficient, since it does not make use of the exposure and IV measured among cases. Under certain assumptions, it may be possible to improve the efficiency of the first stage regression which in turn may lead to a more efficient second stage estimator of the treatment effect. This can be achieved by using all available information on both cases and controls, and by adjusting for case-control status in estimating the first stage regression model. For instance, for continuous A, one may modify the first stage regression and instead estimate:

$$E(A|Z, Y) = \alpha_0 + \alpha_1 Z + \alpha_2 Y,$$

which involves adjusting for ascertainment by directly conditioning on case-control status in the regression model. Under the rare disease assumption, the above model would in principle recover an unbiased estimator $(\widetilde{\alpha}_0, \widetilde{\alpha}_1)$ of (α_0, α_1) provided the degree of ascertainment bias (here encoded by a non-null value of α_2) does not vary with Z (Tchetgen Tchetgen et al, 2013, Tchetgen Tchetgen, 2013b). It is important to note that some care is needed in forming the residual used in the second stage logistic regression, which must reflect the residual value for A in the underlying population and is therefore obtained by evaluating the predicted mean of A under the above estimated mean model, after setting $Y = 0$ for both cases and controls (Tchetgen Tchetgen et al, 2013), i.e.

$$\widehat{\Delta} = A - \widehat{E}(A|Z, Y = 0) = A - \widetilde{\alpha}_0 - \widetilde{\alpha}_1 Z$$

Tchetgen Tchetgen (2013b) discusses analogous methodology to account for possible heterogeneity in the degree of selection bias, and similar techniques are also developed in the context of logistic regression for binary A, and are likewise extended to account for case-control ascertainment when Y is not necessarily rare in the population. However, similar to standard inverse probability weighting, which may also be used to account for the sampling design (although it may be relatively inefficient), the sampling fractions for cases and controls must be available to account for sampling conditional on an outcome Y which may not be rare in the target population (Tchetgen Tchetgen, 2013b).

Adjusting for covariates

Here we consider a straightforward generalization to allow for the presence of covariates C such that Z is a valid IV conditional on C but not necessarily so upon marginalizing over any component of C. Assuming standard prospective sampling, in order to incorporate such covariates, it suffices to modify regression models used in the first and second stages, such that in the case of continuous exposure, the first stage regression further adjusts for C, e.g.

$$E(A|Z,C) = \alpha_0 + \alpha_1 Z + \alpha_2 C,$$

and likewise, for binary A, one could specify

$$\operatorname{logit} \Pr(A=1|Z,C) = \alpha_0 + \alpha_1 Z + \alpha_2 C.$$

The second stage regression in the rare outcome situation could also be modified accordingly, e.g.

$$\operatorname{logit} \Pr(Y=1|A,Z) = \beta_0^* + \beta_C' C + \beta_A A + \ (\alpha_1 + \alpha_2 Z)\,\widehat{\Delta}$$

with $\widehat{\Delta}$ the estimated residual $A - \widehat{E}(A|Z,C)$, and analogous adjustments can be made to the risk ratio regression approach recommended for non-rare outcomes, as well as under case-control sampling.

Conclusion

In this note, an alternative framework is proposed to motivate the control function IV approach in the context of binary outcome, with binary treatment. Although emphasis is given to binary treatment, the approach can be modified to handle other types of discrete treatment, without much difficulty. The approach can also be used with a continuous IV without further difficulty. In addition, unlike available formulations of the control function approach, the proposed framework allows for the presence of heterogeneity in the magnitude of selection bias (on the risk ratio scale) with respect to the IV. Such heterogeneity may reflect latent heterogeneity in the degree of association of the IV with the treatment, the presence of which cannot be ruled out with certainty in practice. Ignoring such heterogeneity when present may invalidate the commonly used control function approach, and therefore the proposed framework is recommended for routine use as a more robust alternative strategy for IV estimation in the context of binary outcome and binary treatment.

Appendix

Proof of Result 1: *Note that*

$$\begin{aligned}\Pr\left(Y=1|A,Z\right) &= E\left[\exp\left(\beta_0+\beta_A A+U\right)|A,Z\right] \\ &= \exp\left(\beta_0+\beta_A A\right)E\left[\exp\left(U\right)|A,Z\right] \\ &= \exp\left(\beta_0+\beta_A A+\log E\left[\exp\left(\delta\right)\right]\right) \\ &\times \exp\left\{E\left(U|A,Z\right)\right\}.\end{aligned}$$

Further note that

$$
\begin{aligned}
E\left(U|A,Z\right) &= E\left(U|A,Z\right) - E\left(U|A=0,Z\right) \\
&\quad - \int \left\{E\left(U|a,Z\right) - E\left(U|A=0,Z\right)\right\} dF\left(a|Z\right) \\
&\quad + E\left(U|Z\right) \\
&= \omega_1 A + \omega_2 AZ - \alpha_1 E\left(A|Z\right) - \alpha_2 E\left(A|Z\right) Z \\
&\quad + E\left(U\right)
\end{aligned}
$$

where

$$\omega_1 = E\left(U|A=1,Z=0\right) - E\left(U|A=0,Z=0\right)$$

and

$$
\begin{aligned}
\omega_2 &= E\left(U|A=1,Z=1\right) - E\left(U|A=0,Z=1\right) \\
&\quad - \left\{E\left(U|A=1,Z=0\right) - E\left(U|A=0,Z=0\right)\right\}
\end{aligned}
$$

therefore

$$\Pr(Y=1|A,Z) = \exp\left\{\beta_0^* + \beta_A A + \ \left(\omega_1 + \omega_2 Z\right)\Delta\right\}$$

where

$$\beta_0^* = \beta_0 + \log E\left[\exp\left(\delta\right)\right] + E\left(U\right)$$

proving the result.

References

[1] Angrist JD. Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. J Bus Econ Stat. 2001;19(1):2–28.

[2] Basmann RL. A generalized classical method of linear estimation of coefficients in a structural equation. Econometrica. 1957;25(1):77–83.Davey Smith G, Ebrahim S. ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32(1):1–22.

[3] Blundell RW, Powell JL. Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont M, Hansen LP, Turnovsky SJ, eds. Advances in Economics and Econometrics: Theory and Applications. 8th World Congress of the Econometric Society. Cambridge, United Kingdom: Cambridge University Press; 2003:312–357.

[4] Didelez V, Meng S, Sheehan NA. Assumptions of IV methods for observational epidemiology. Stat Sci. 2010;25(1):22–40.

[5] Garen J. The returns to schooling: a selectivity bias approach with a continuous choice variable. Econometrica. 1984;52(5):1199–1218.

[6] Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol. 2000;29(4):722–729.

[7] Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist’s dream? Epidemiology. 2006;17(4):360–372.

[8] Lawlor DA, Harbord RM, Sterne JA, et al. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med. 2008;27(8):1133–1163.

[9] Mullahy J. Instrumental-variable estimation of count data models: applications to models of cigarette smoking behaviour. Rev Econ Stat. 1997;79(4):568–593.

[10] Nagelkerke N, Fidler V, Bernsen R, et al. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. Stat Med. 2000;19(14): 1849–1864.

[11] Palmer, TM, et al. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. American journal of epidemiology 173.12 (2011): 1392-1403.

[12] Tchetgen Tchetgen E.J. (2013) Estimation of risk ratios in cohort studies with a common outcome: a simple and efficient two-stage approach.Int J Biostat. 2013 May 7;9(2):251-64. doi: 10.1515/ijb-2013-0007.

[13] Tchetgen Tchetgen, EJ, Walter S, Glymour MM:Building an evidence base for Mendelian Randomization studies (2013). IJE 42 (1), 328-331.

[14] Tchetgen Tchetgen EJ. A general regression framework for a secondary outcome in case-control studies. (2013b) Biostatistics. 15 (1): 117-128.

[15] Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. J Health Econ. 2008;27(3):531–543.

[16] Theil H. Repeated Least Squares Applied to Complete Equation Systems. The Hague, the Netherlands: Central Planning Bureau; 1953.

[17] Vansteelandt, S, et al. On instrumental variables estimation of causal odds ratios. Statistical Science 26.3 (2011): 403-422.

[18] Wacholder S.(1986) Binomial regression in GLIM: estimating risk ratios and risk differences. Am J Epidemiol;123:174 –184.

[19] Wooldridge JM. On two stage least squares estimation of the average treatment effect in a random coefficient model. Econ Lett. 1997;56(2):129–133.

[20] Wooldridge JM. Econometric Analysis of Cross Section and Panel Data. 2nd ed. Cambridge, United Kingdom: MIT Press; 2002.

[21] Zou GY.(2004) A modified Poisson regression approach to prospective studies with binary data. Am J Epidemiol.;159:702–706.