The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers
www.cesifo.org/wp

Generalized Propensity Scores for Multiple Continuous Treatment Variables

Peter H. Egger
Maximilian von Ehrlich

CESIFO WORKING PAPER NO. 4074
CATEGORY 12: EMPIRICAL AND THEORETICAL METHODS
JANUARY 2013

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 4074*

Generalized Propensity Scores for Multiple Continuous Treatment Variables

Abstract

This paper illustrates that the generalized propensity score method can easily be applied with multiple continuous endogenous treatment variables. Consistency proofs carry over straightforwardly to this general case, and the approach is shown to work well in finite samples with various data-generating processes and up to five continuous endogenous treatment variables.

JEL-Code: C140, C210.

Keywords: generalized propensity score estimation, multiple treatments, continuous endogenous treatments.

Peter H. Egger
ETH Zurich
Weinbergstrasse 35
Switzerland – 8092 Zurich
egger@kof.ethz.ch

Maximilian von Ehrlich
ETH Zurich
Weinbergstrasse 35
Switzerland – 8092 Zurich
vonehrlich@kof.ethz.ch

December 10, 2012

1 Introduction

Imbens (2000), Lechner (2001), Hirano and Imbens (2004) and Imai and van Dyk (2004) should be credited with an invaluable generalization of the seminal propensity score method by Rosenbaum and Rubin (1983) to the case of multivalued and, in particular, continuous endogenous treatments. The idea is to utilize the propensity score as a scalar-valued control function which is based on observable characteristics in the data in order to remove the endogeneity bias in average effects of endogenous treatments on outcome of interest. For instance, Hirano and Imbens (2004) show that this works well with continuous treatments when utilizing the residuals from an estimated continuous-treatment-generating econometric model to generate a generalized propensity score which assumes normality and is used in a flexible (e.g., polynomial) functional form as a control function in the outcome-generating model.

This short paper follows up on this idea. It shows that the approach of Hirano and Imbens (2004) is as well applicable with multiple correlated continuous endogenous treatments as it is with single treatment variables. We illustrate that consistency proofs are straightforward and the small sample properties of the estimator are similar between single-treatment and multiple-treatment frameworks. The next section introduces the notation and outlines the approach for multiple treatments with the large-sample properties being relegated to the Appendix, and Section 3 summarizes the small sample properties for three data-generating processes and one to five endogenous treatments.

2 Generalized propensity scores for multiple treatments

We wish to nonparametrically estimate the average treatment effect function (ATE) of M continuous, endogenous treatments which are indexed by $m = 1, ..., M$ on outcome Y_i of cross-sectional units $i = 1, ..., N$. There are three treatment concepts. First, denote the mth set of *potential treatment levels* by $\mathfrak{T}_m \in [\underline{t}_m, \bar{t}_m]$, where $\underline{t}_m$ and $\bar{t}_m$ are the corresponding lower and upper bounds, respectively. Second, denote particular levels of potential treatment in the interval $[\underline{t}_m, \bar{t}_m]$ by $t_m \in \mathfrak{T}_m$. Finally, refer to *actual treatment levels* for unit i by T_{mi}, and the combinations of potential and realized treatments in M dimensions by vectors $t = t_1, ..., t_M$ and $T_i = (T_{1i}, \ldots, T_{Mi})$, respectively.

Postulate outcome Y_i as a flexible function of T_{mi} as $Y_i(T_i) = f(T_{1i}, ..., T_{Mi})$ and of potential treatments as $Y_i(t) = Y_i(t_1, ..., t_M)$. The latter may be referred to as

the unit-level dose-response function, whose average across units i is the average dose-response function, $\mu(t) \equiv E[Y_i(t)]$. Specify T_{mi} as a function of nonstochastic regressor vector X_{mi} which may potentially be correlated with Y_i:

$$T_{mi} = f(X_{mi}, \delta_m) + \varepsilon_{mi}, \tag{1}$$

where δ_m is an unknown parameter vector and ε_{mi} is a stochastic term which is uncorrelated with both X_{mi} and Y_i. Define the joint matrix of nonstochastic regressors (instruments) in the system by Z_i which contains at least $X_{1i} \cup ... \cup X_{Mi}$ and possibly also interactive terms of elements of the individual vectors X_{mi}, so that we may formulate a reduced-form specification for all M treatments as

$$T_{mi} = \underbrace{f(Z_i, \gamma_m)}_{=\overline{T}_{mi}} + \nu_{mi}, \tag{2}$$

where γ_m is an unknown parameter vector and ν_{mi} is a stochastic term which is uncorrelated with both Z_i and Y_i.

For identification, we have to assume weak unconfoundedness as stated in Rosenbaum and Rubin (1983) for the binary propensity score and in Hirano and Imbens (2004) and Imai and van Dyk (2004) for the generalized propensity score with a single, multi-valued (continuous) treatment.

Assumption (Weak Unconfoundedness)

$$Y_i(t) \perp T_{1i}, ..., T_{Mi} \,|Z_i \; \forall t_1 \in \mathfrak{T}_1, ..., t_M \in \mathfrak{T}_m.$$

Hence, the potential outcome $Y_i(t)$ is conditionally independent of treatment status T_m. The generalized propensity score in the M-dimensional continuous treatment is specified as follows.

Definition (Generalized Propensity Score)
Denote any possible vector of covariates determining treatment by z and define the M-variate conditional joint density of $t_1, ..., t_M$ given z as

$$g(t, z) = f_{T_i|Z_i}(t|z).$$

Then, the generalized propensity score (GPS) is defined as

$$G_i = g(T_i, Z_i), \quad Z_i \perp 1\{T_{mi} = t_m \forall m = 1, ..., M\} \,|g(t, Z_i)\,.$$

Hence, the probability of $T_i = (T_{1i}, ..., T_{Mi})$ being equal to some potential treatment combination t is independent of the covariates in Z_i once we condition on the GPS.

Accordingly, treatment status is independent of outcome conditional on the GPS once the above assumption is met. For identification, this implies that under weak unconfoundedness conditioning on (some function of) the scalar-valued G_i instead of (some function of) all elements in Z_i is sufficient to remove the selection bias in the unconditional impact of all treatments on outcome.

Let us denote the $N \times 1$ GPS vector by $G = (G_i)$, the $N \times M$ matrix of demeaned treatments by $\widetilde{T} = (\nu_{mi})$ with $\nu_{mi} = T_{mi} - \overline{T}_{mi}$ where $\overline{T}_{mi}$ is the (conditional) mean of T_{mi}, and the $M \times M$ symmetric and positive definite variance-covariance matrix of treatments by $\Sigma = Cov[\nu_m, \nu_{m'}]$ where $\nu_m = (\nu_{mi})$ and $\nu_{m'} = (\nu_{m'i})$ denote two $N \times 1$ residual vectors for treatments m and m'. Then, $G = (G_i)$, and its estimated counterpart $\widehat{G} = (\widehat{G}_i)$ based on the multivariate normal are given by

$$G = \frac{1}{(2\pi)^{M/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}\widetilde{T}'\Sigma^{-1}\widetilde{T}\Big), \;\; \widehat{G} = \frac{1}{(2\pi)^{M/2}|\widehat{\Sigma}|^{1/2}} \exp\Big(-\frac{1}{2}\widehat{\widetilde{T}}'\widehat{\Sigma}^{-1}\widehat{\widetilde{T}}\Big). \quad (3)$$

The estimated $\widehat{G} = (\widehat{G}_i)$ can then be used in a flexible control function to reduce (if not remove) the endogeneity bias of the estimated average treatment effect in the model determining outcome Y_i. The Appendix proves consistency of this approach and the next section illustrates its small-sample performance.

3 Monte Carlo set-up and simulation results

Consider the treatment-generating process

$$T_{mi} = \overline{T}_{mi} + \nu_{mi} = X_{mi}\beta_m + \nu_{mi}$$

where $X_{mi} \sim i.i.d.N(0,1)$ and $\beta_m = 5$ for each $m = 1, \ldots, M$, and

$$\nu_i = [\nu_{1i}, \ldots, \nu_{Mi}]' \sim i.i.d.N(0, \Sigma),$$

where all diagonal elements of Σ are assumed to be unity and all off-diagonal elements are assumed to be 0.25, for simplicity. Let us specify $L_i = [T_{1i}, \ldots, T_{Mi}, T_{1i}^2, \ldots, T_{Mi}^2, T_{1i}^3, \ldots, T_{Mi}^3]$, $H_i = [X_{1i}, \ldots, X_{Mi}, X_{1i}^2, \ldots, X_{Mi}^2]$, $\Xi_i = [T_{1i}X_{1i}, \ldots, T_{Mi}X_{Mi}]$, and $\Gamma_i = [G_i^1, \ldots, G_i^3, G_iT_{1i}, \ldots, G_i^3T_{Mi}^3]$ to formulate three considered processes for outcome:

$$Y_i^A = L_i\alpha_1 + \Gamma_i\alpha_2 + u_i \quad (4)$$

$$Y_i^B = L_i\alpha_1 + H_i\alpha_3 + u_i \quad (5)$$

$$Y_i^C = L_i\alpha_1 + H_i\alpha_3 + \Xi_i\alpha_4 + u_i \quad (6)$$

where $u_i \sim i.i.d.N(0,s)$ and s scales the variance of u_i. We set all elements in the vectors α_k for $k = 1,\ldots,4$ to unity. In the counterfactual situation, we raise T_{1i} to $T_{1i}^c = T_{1i} + 0.01$ at Z_i without loss of generality. This changes ν_{1i} and G_i and, in turn, it changes Y_i^j to Y_i^{jc} for $j \in \{A,B,C\}$. Let us denote the true average treatment effect of such a change in T_{1i} on outcome Y_i^j by $ATE_1^j = Y_i^{jc}/Y_i^j - 1$. We aim at estimating the latter by OLS, assuming linearity and mean independence ($\widehat{ATE}_{1OLS}^j = \widehat{\alpha}_{1,1}$) and, alternatively, by the GPS estimator ($\widehat{ATE}_{1GPS}^j$). Specifically, after defining $\widehat{\Upsilon}_i = [\widehat{G}_i^1,\ldots,\widehat{G}_i^3,\widehat{G}_iT_{1i},\ldots,\widehat{G}_i^5T_{1i}^5]$, $\widehat{\nu}_{mi} = T_{mi} - \widehat{T}_{mi}$, and $Z_i = [X_{i1},\ldots,X_{iM}]$, the GPS-based estimates are obtained as follows:[1]

$$\text{First stage: } \widehat{T}_{mi} \;=\; Z_i\widehat{\vartheta}_m; \text{Second stage: } \widehat{Y}_i^j = [L_i, \widehat{\Upsilon}_i]\widehat{\varphi}^j, \tag{7}$$

where $\widehat{\vartheta}_m$ and $\widehat{\varphi}^j$ are estimated conformable parameter vectors. Notice that there are two approximation errors in (7). First, (7) ignores the exclusion restrictions (that T_{mi} only depends on X_{mi} only) and, second, the functional form in which Y_i^j depends on T_{1i} in (7) is different from the true processes in (4)-(6).

We consider cases of $M \in \{1,\ldots,5\}$, of $N \in [1,200; 2,400; 4,800]$ observations in the data, and of two configurations for the scaling factor of the variances in the second-stage models, $s \in \{1,10\}$. Altogether, this gives $5 \cdot 3 \cdot 2 = 30$ experiments for which we do $2,000$ Monte Carlo runs each. The results for the average bias and root mean squared error (RMSE) across all runs and observations i within an experiment are summarized in Table 1.

Table 1 suggests the following conclusions. First of all, OLS assuming ATE linearity is always dramatically biased and exhibits a large RMSE. This is not surprising, given the assumed high degree of ATE nonlinearity and endogeneity in the data-generating process both of which are ignored by OLS across all Models A-C. Obviously, bias as well as RMSE are always higher with more noise in the outcome process, i.e., for $s = 10$ compared to $s = 1$. Moreover, OLS tends to perform weakest for Model C due to ignoring the interactive terms in Ξ.

Second, among Models A-C, the GPS approach works relatively best for Model A. The reason is simply that the control function is perfectly specified in expected value in that model. Hence, the only error accrues to pure stochastics. Clearly, this is not what one will encounter empirically. In Models B and C there is a polynomial approximation error about the control function, which leads to higher bias than in Model A.[2] Third, as expected, the bias declines as N gets bigger (to the extent that

[1] Alternatively, we considered approximations, where $\widehat{\Upsilon}_i$ was based on a (less flexible) form with $P = 3$ and a (more flexible) form with $P = 10$. The results are summarized in an online appendix. Naturally, the bias declines with the degree of flexibility of the control function.

[2] Empirically, this could be further reduced by a search algorithm about the functional form of the control function, e.g., based on information criteria.

there is small sample bias on top of functional form bias of the control function). Naturally, the RMSE declines as N increases. Moreover, the RMSE rises as the signal-to-noise ratio declines in the outcome equation with much less impact on the bias.

Finally, the extent of bias and RMSE is largely invariant to an increase in M. Hence, the GPS approach works as well with multiple treatments as with a single treatment.

References

Hirano, K., Imbens, G. 2004. The propensity score with continuous treatment. In A. Gelman and X.-L. Meng (eds.), Applied Bayesian Modelling and Causal Inference from Missing Data Perspectives. New York: Wiley, 73-84.

Imai, K., van Dyk, D.A. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association 99(467), 854-66.

Imbens, G. 2000. The role of the propensity score in estimating dose-response functions. Biometrika 87(3), 706-710.

Lechner, M. 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In M. Lechner and F. Pfeiffer (eds.), Econometric Evaluation of Active Labour Market Policies in Europe. Heidelberg: Physica Verlag, 43-58.

Rosenbaum, P.R., Rubin, D.B. 1983. The central role of the propensity score in observational studies for causal effects." Biometrika, 70(1), 41-55.

Appendix

For ease of notation, we suppress subscripts for individuals in this section.

Theorem 1 *If the assignment to treatment is weakly unconfounded given pre-treatment covariates Z, then for every potential multiple treatment level t*

$$f_T\left(t|g(t,Z),Y(t)\right) = f_T\left(t|g(t,Z)\right) \tag{8}$$

Proof 1 *Denote the conditional probability distribution for Z by $F_Z(z|.)$ and the conditional densities of $T=(T_1,...,T_M)$ by $f_T(t|.)$. Weak unconfoundedness means*

$$f_T\left(t|z,g(t,Z),Y(t)\right) = f_T\left(t|z\right) = g(t,z).$$

This is the case, since $f_T(t|Z,g(t,Z),Y(t)) = f_T(t|Z,g(t,Z)) = f_T(t|g(t,Z)) = g(t,Z)$, so that (8) holds under weak unconfoundedness whereby treatments and outcome are mutually independent conditional on the GPS for multiple treatments.

Theorem 2 *Denote the conditional expectation of outcome by $\eta(t,g)$ and the average dose response function by $\mu(t)$. Referring to element-wise equality $T_m = t_m$ for all $m = 1, ..., M$ by $T = t$, under weak unconfoundedness we have (a) $\eta(t,g) = E[Y(t)|g(t,Z) = g] = E[Y|T = t, G = g]$ and (b) $\mu(t) = E[\eta(t, g(t,Z)]$.*

Proof 2 *Denote the conditional density of $Y(t) = y$ conditional on $T = t$ and $g(t,Z) = g$ by $f_{Y(t)}(y|t,g)$. Using Bayes' rule and Theorem 1, we obtain*

$$f_{Y(t)}(y|t,g) \quad = \quad \frac{f_T(t|Y(t) = y, g(t,Z) = g) f_{Y(t)}(y|g)}{f_T(t|g(t,Z) = g)} = f_{Y(t)}(y|g).$$

With $E[Y(t)|T = t, g(t,Z) = g] = E[Y(t)|g(t,Z)]$,

$$E[Y|T = t, G = g] \quad = \quad E[Y(t)|T = t, g(T,Z) = g] = E[Y(t)|g(t,Z) = g] = \eta(t,g), \tag{9}$$

which proves Part (a). Estimating $E[Y|T = t, G = g]$ yields the parameters needed for calculating the dose response function. Part (b) follows from (9) together with the law of iterated expections:

$$E[\eta(t, g(t,Z)] = E[E[Y(t)|g(t,Z) = g]] = E[Y(t)].$$