Dummy Endogenous Variables in Weakly Separable Multiple Index Models without Monotonicity*

Songnian Chen — HKUST
Shakeeb Khan — Boston College
Xun Tang — Rice University

April 1, 2020

Abstract

We study the identification and estimation of treatment effect parameters in weakly separable models. In their seminal work, Vytlacil and Yildiz (2007) showed how to identify and estimate the average treatment effect of a dummy endogenous variable when the outcome is weakly separable in a single index. Their identification result builds on a monotonicity condition with respect to this single index. In comparison, we consider similar weakly separable models with multiple indices, and relax the monotonicity condition for identification. Unlike Vytlacil and Yildiz (2007), we exploit the full information in the distribution of the outcome variable, instead of just its mean. Indeed, when the outcome distribution function is more informative than the mean, our method is applicable to more general settings than theirs; in particular we do not rely on their monotonicity assumption and at the same time we also allow for multiple indices. To illustrate the advantage of our approach, we provide examples of models where our approach can identify parameters of interest whereas existing methods would fail. These examples include models with multiple unobserved disturbance terms such as the Roy model and multinomial choice models with dummy endogenous variables, as well as potential outcome models with endogenous random coefficients. Our method is easy to implement and can be applied to a wide class of models. We establish standard asymptotic properties such as consistency and asymptotic normality.

JEL Classification: C14, C31, C35

Key Words Weak Separability, Treatment Effects, Monotonicity, Endogeneity

*We are grateful to Jeremy Fox, Sukjin Han, Arthur Lewbel, Elie Tamer, Ed Vytlacil, Haiqing Xu, and participants at the 2018 West Indies Economic Conference, the 2019 SUFE econometrics meetings, and the 2020 Texas Camp Econometrics for helpful comments and suggestions.

1 Introduction

Consider a weakly separable model with a binary endogenous variable:

$$Y = g(v_1(X,D), v_2(X,D), ...v_J(X,D), \varepsilon) \quad (1.1)$$

$$D = 1\{\theta(Z) - U > 0\} \quad (1.2)$$

where $(v_1(X,D), v_2(X,D), ...v_J(X,D)) \equiv v(X,D)$ is a J-vector of unknown linear or nonlinear indices in the outcome equation (1.1) and D is a binary endogenous variable defined by the selection equation (1.2). Here $X \in \mathbb{R}^{d_x}$ and $Z \in \mathbb{R}^{d_z}$ are vectors of observable exogenous variables, which may have overlapping elements. Similar to Vytlacil and Yildiz (2007) we require exclusion restrictions that there is some element in Z excluded from X, and that we can vary X after conditioning on $\theta(Z)$. In the system of equations above, U is the unobservable random variable normalized to follow the uniform distribution $U(0,1)$ and the error term ε in the outcome equation is allowed to be a random vector. We assume (X,Z) are independent of (ε, U). Note that we allow $v(X,D)$ to be a vector of multiple indices, whereas the method in Vytlacil and Yildiz (2007) can only be applied when it is a single index.

Since Vytlacil and Yildiz (2007), other important work has considered identification and estimation of related models, but under alternative conditions. Examples with binary endogenous variables include Han and Vytlacil (2017), Vuong and Xu (2017), Lewbel, Jacho-Chavez, and Encarnciono (2016), Khan, Maurel, and Zhang (2019). Work for models when the endogenous variable is continuous includes Imbens and Newey (2009), D'Haultfoeuille and Fevrier (2015) and Torgovitsky (2015). Feng (2020) shows how to identify nonseparable triangular models where the endogenous variable is discrete and has larger support than the instrument variable.[1]

As in the conventional framework, two potential outcomes Y_1 and Y_0 satisfy

$$Y_D = g(v(X,D), \epsilon) \text{ for } D = 0, 1.$$

We only observe (Y, D, X, Z), where $Y = DY_1 + (1-D)Y_0$. In this model, as in Vytlacil and Yildiz (2007), we do not impose parametric distribution on the error term or a linear index structure. Vytlacil and Yildiz (2007) assumes that $v(X,D) \in \mathbb{R}$ is a single index, and

$$E[g(v,\varepsilon)|U=u] \text{ is strictly increasing in } v \in \mathbb{R} \text{ for all } u. \quad (1.3)$$

[1] All these papers focus on point identification. For partial identification of a model with a binary outcome, see Shaikh and Vytlacil (2011) and Mourifié (2015).

Unlike Vytlacil and Yildiz (2007), we do not impose any monotonicity structure. That is because our approach exploits the full information in the distribution of the outcome variable, instead of just its mean. Indeed, when the outcome distribution function is more informative than the mean, our method is applicable to more general settings than theirs; in particular we do not rely on their monotonicity assumption and at the same time we also allow for multiple indices.

In Sections 3 and 4 we provide some examples in which such a monotonicity condition fails, but the average effect of the binary endogenous variable is still identified. In addition, we allow for a weakly separable model with multiple indices, that is, $v(X, D) = (v_1(X, D), v_2(X, D), ..., v_J(X, D)) \in \mathbb{R}^J$.

We consider the identification and estimation of the average treatment effect of D on Y, $E(Y_1|X \in A)$, $E(Y_0|X \in A)$ and $E(Y_1 - Y_0|X \in A)$, for some set A, without the aforementioned monotonicity. Indeed, for the case with multiple indices $v(X, D) \in \mathbb{R}^J$, the monotonicity condition is no longer well defined.

Vuong and Xu (2017) established nonparametric identification of individual treatment effects in a fully nonseparable model that includes a binary endogenous regressor, without the nonlinear index structure. They assume ε is a scalar and g is strictly increasing in ε. In their setting, monotonicity in the outcome equation provides the identifying restriction to extrapolate information from local treatment effects to population treatment effects.

2 Identification

Generally speaking, our identification strategy will be based on the notion of *matching*[2]. Consider the identification of $E(Y_1|X = x)$ for some $x \in S_1$, where S_d denotes the support of X given $D = d \in \{0, 1\}$. Note that because $(\varepsilon, U) \perp (X, Z)$,

$$\begin{aligned} E(Y_1|X = x) &= E(Y_1|X = x, Z = z) \\ &= E(DY_1|X = x, Z = z) + E[(1 - D)Y_1|X = x, Z = z] \\ &= P(z)E(Y|D = 1, X = x, Z = z) + [1 - P(z)]E(Y_1|D = 0, X = x, Z = z) \quad (2.1) \end{aligned}$$

[2]See Ahn and Powell (1993), Chen, Khan, and Tang (2016), Vytlacil and Yildiz (2007), and more recently Auerbach (2019) for examples of papers that attain identification through matching.

where $P(z) \equiv E(D|Z=z)$. The only term that is not directly identifiable on the right-hand side of (2.1) is

$$E(Y_1|D=0, X=x, Z=z) = E[g(v(x,1),\varepsilon)|U \geq P(z)].$$

The main idea behind our approach follows that of Vytlacil and Yildiz (2007), which is to find some $\tilde{x} \in S_0$ such that

$$v(x,1) = v(\tilde{x},0) \tag{2.2}$$

so that

$$\begin{aligned} E(Y|D=0, X=\tilde{x}, Z=z) = &\quad E(Y_0|D=0, X=\tilde{x}, Z=z) \\ = &\quad E(g(v(\tilde{x},0),\varepsilon)|U \geq P(z)) = E(g(v(x,1),\varepsilon)|U \geq P(z)). \end{aligned}$$

Unlike Vytlacil and Yildiz (2007), we utilize the full distribution of Y (rather than its first moment) while searching for such pairs of $(x, \tilde{x})$ in (2.2). This allows us to relax the single-index and monotonicity conditions in Vytlacil and Yildiz (2007).

For any p on the support of $P(Z)$ given $X=x$, and for all y define

$$\begin{aligned} h_1^*(x,y,p) &= E(D1\{Y \leq y\}|X=x, P(Z)=p) \\ &= E[1\{U < P(Z)\} 1\{g(v(X,1),\varepsilon) \leq y\}|X=x, P(Z)=p] \\ &= \int_0^p F_{g|u}(y; v(x,1))du, \end{aligned} \tag{2.3}$$

where

$$F_{g|u}(y; v(x,d)) \equiv E[1\{g(v(x,d),\varepsilon) \leq y\}|U=u]$$

with $v(x,d)$ being a realized index at $X=x$ and the expectation in the definition of $F_{g|u}$ is with respect to the distribution of ε given $U=u$. The last equality in (2.3) holds because of independence between (ε, U) and (X, Z). By construction, $h_1^*(x,y,p)$ is directly identified from the joint distribution of (D, Y, X, Z) in the data-generating process. Furthermore, for any pair $p_1 > p_2$, define:

$$h_1(x,y,p_1,p_2) \equiv h_1^*(x,y,p_1) - h_1^*(x,y,p_2) = \int_{p_2}^{p_1} F_{g|u}(y; v(x,1))du.$$

Likewise, define

$$\begin{aligned} h_0^*(x,y,p) &= E((1-D)1\{Y \leq y\}|X=x, P(Z)=p) \\ &= E[1\{U \geq P(Z)\} 1\{g(v(X,0),\varepsilon) \leq y\}|X=x, P(Z)=p] \\ &= \int_p^1 F_{g|u}(y; v(x,0))du. \end{aligned}$$

and let

$$h_0(x, y, p_1, p_2) \equiv h_0^*(x, y, p_2) - h_0^*(x, y, p_1) = \int_{p_2}^{p_1} F_{g|u}(y; v(x, 0))du.$$

Let P_x denote the support of $P(Z)$ given $X = x$. It can be shown that for any $x \in S_1$ and $\tilde{x} \in S_0$, and any y,

$$h_1(x, y, p, p') = h_0(\tilde{x}, y, p, p') \text{ for all } p > p' \text{ on } P_x \cap P_{\tilde{x}}. \tag{2.4}$$

if and only if

$$F_{g|p}(y; v(x, 1)) = F_{g|p}(y; v(\tilde{x}, 0)) \text{ for all } p \in P_x \cap P_{\tilde{x}}. \tag{2.5}$$

Sufficiency is immediate from the definition of h_1 and h_0. To see necessity, note that for all $p > p'$ on $P_x \cap P_{\tilde{x}}$,

$$\left.\frac{\partial h_1(x, y, \tilde{p}, p')}{\partial \tilde{p}}\right|_{\tilde{p}=p} = \left.\frac{\partial h_1^*(x, y, \tilde{p})}{\partial \tilde{p}}\right|_{\tilde{p}=p} = F_{g|p}(y; v(x_1, 1))$$

and

$$\left.\frac{\partial h_0(\tilde{x}, y, \tilde{p}, p')}{\partial \tilde{p}}\right|_{\tilde{p}=p} = -\left.\frac{\partial}{\partial \tilde{p}} h_0^*(\tilde{x}, y, \tilde{p})\right|_{\tilde{p}=p} = F_{g|p}(y; v(\tilde{x}, 0)).$$

Thus (2.4) and (2.5) are equivalent.

We collect the assumptions for identification as follows:

ASSUMPTION A-1: The distribution of U is absolutely continuous with respect to Lebesgue measure.

ASSUMPTION A-2: The random vectors (U, ε) and (X, Z) are independent.

ASSUMPTION A-3: The random variable $g(v(X, 1), \varepsilon)$ and $g(v(X, 0), \varepsilon)$ have finite first moments conditional on $U = u$ for all $u \in [0, 1]$..

ASSUMPTION A-4: For any $(x, \tilde{x}) \in S_1 \times S_0$, $F_{g|p}(y; v(x, 1)) = F_{g|p}(y; v(x, 0))$ holds for all y and $p \in P_x \cap P_{\tilde{x}}$ if and only if $v(x, 1) = v(\tilde{x}, 0)$.

ASSUMPTION A-5: $\Pr(X \in S_1) > 0$ and $\Pr(X \in S_0) > 0$.

Note that A-4 is weaker than Assumption 4 in Vytlacil and Yildiz (2007). Specifically, to identify pairs $(x, \tilde{x})$ with $v(x, 1) = v(\tilde{x}, 0)$, Vytlacil and Yildiz (2007) relies on the assumption that $v(X, D) \in \mathbb{R}$ is a single index and that for any $(x, \tilde{x}) \in (S_1 \times S_0)$, $E[g(v(x, 1), \varepsilon)|U = u] = E[g(v(\tilde{x}, 0), \varepsilon)|U = u]$ if and only if $v(x, 1) = v(\tilde{x}, 0)$. There are two shortcomings with this approach. First, it requires the condition (Assumption 4) that $E[g(v(x, d), \varepsilon)|U = p]$ is a strictly monotonic function of $v(x, d)$. Second, when $v(x, d)$ is a vector of multiple indices instead of a single index, their approach breaks down. In comparison, we achieve the same purpose by matching conditional distributions $F_{g|p}(\cdot; v(x, 1))$ and $F_{g|p}(\cdot; v(\tilde{x}, 0))$. As we show in Section 3, in several important applications, the outcome Y is either discrete (e.g. multinomial choices), or multi-dimensional with both discrete and continuous components (e.g., potential outcomes determined by a Roy model). In either cases, the latent index function $v(.)$ is vector-valued and the monotonicity condition in Vytlacil and Yildiz (2007) is not satisfied.

3 Examples

We now present several examples in which the latent indices are multi-dimensional. In the first and third example, the monotonicity condition in Vytlacil and Yildiz (2007) is not satisfied; in the second example, the identification requires a generalization of the monotonicity condition into an invertibility condition in higher dimensions.

Example 1. (Heteroskaedastic shocks in outcome) Consider a triangular system where a continuous outcome is determined by double indices $v(X, D) \equiv (v_1(X, D), v_2(X, D))$:

$$Y = g(v(X, D), \varepsilon) = v_1(X, D) + v_2(X, D)\varepsilon \text{ for } D \in \{0, 1\}.$$

The selection equation determining the actual treatment is the same as (1.2). In this case the concept of monotonicity in $v \in \mathbb{R}^2$ is not well-defined, so the procedure proposed in Vytlacil and Yildiz (2007) is not suitable here[3]. Nevertheless, we can apply the method in Section 2 to identify the average treatment effect by using the *distribution* of outcome to find pairs of x and $\tilde{x}$ such that $v(x, 1) = v(\tilde{x}, 0)$. Assume the range of $v_2(\cdot)$ is positive. To see the necessity in Assumption A4, note that

$$\begin{aligned} F_{g|u}(y; v(x, d)) &= E[v_1(x, d) + v_2(x, d)\varepsilon \leq y|U = u] \\ &= F_{\varepsilon|u}\left(\frac{y - v_1(x, d)}{v_2(x, d)}\right) \end{aligned}$$

[3]For this particular design, the approach proposed in Vuong and Xu (2017) should be valid. But it will not be for a slightly modified model, such as $Y = v_1(X, D) + (e_2 + v_2(X, D) * e_1)$, whereas ours will be.

for $d = 0, 1$. If the CDF of ε is increasing over $\mathbb{R}$, then for all y and $x \in S_1$ and $\tilde{x} \in S_0$,

$$F_{g|u}(y; v(x, 1)) = F_{g|u}(y; v(\tilde{x}, 0))$$

if and only if

$$\frac{y - v_1(x, 1)}{v_2(x, 1)} = \frac{y - v_1(\tilde{x}, 0)}{v_2(\tilde{x}, 0)}.$$

Differentiating with respect to y yields

$$v_2(x, 1) = v_2(\tilde{x}, 0)$$

which in turn implies

$$v_1(x, 1) = v_1(\tilde{x}, 0).$$

The sufficiency in Assumption A-4 is straight-forward.

Example 2. (Multinomial potential outcome) Consider a triangular system where the outcome is multinomial. The multinomial response model has a long and rich history in both applied and theoretical econometrics. Recent examples in the semiparametric literature include Lee (1995), Ahn, Powell, Ichimura, and Ruud (2017), Shi, Shum, and Song (2018), Pakes and Porter (2014), Khan, Ouyang, and Tamer (2019). But unlike the work here, none of those papers allow for dummy endogenous variables or potential outcomes.

$$Y = g(v(X, D), \varepsilon) = \arg \max_{j=0,1,\ldots,J} y^*_{j,D}$$

where

$$y^*_{j,D} = v_j(X, D) + \varepsilon_j \text{ for } j = 1, 2, ..., J;\ y^*_{0,D} = 0.$$

In this case the index $v \equiv (v_j)_{j \leq J}$ and the errors $\varepsilon \equiv (\varepsilon_j)_{j \leq J}$ are both J-dimensional. The selection equation that determines D is the same as (1.2). In this case, we can replace $1\{Y \leq y\}$ by $1\{Y = y\}$ in the definition of h_1, h_0, h^*_1, h^*_0 and $F_{g|u}(\cdot; v)$. Then for $d = 0, 1$ and $j \leq J$,

$$\begin{aligned} F_{g|u}(j; v(x, d)) &\equiv E[1\{g(v(x, d), \varepsilon) = j\}|U = u] \\ &= \Pr\{v_j(x, d) + \varepsilon_j \geq v_{j'}(x, d) + \varepsilon_{j'}\ \forall j' \leq J \mid U = u\}. \end{aligned}$$

By Ruud (2000) and Ahn, Powell, Ichimura, and Ruud (2017), the mapping from $v \in \mathbb{R}^J$ to $(F_{g|u}(j; v) : j \leq J) \in \mathbb{R}^J$ is smooth and invertible provided that $\varepsilon \in \mathbb{R}^J$ has non-negative density everywhere. This implies Assumption A-4.

Example 3. **(Potential outcome from the Roy model)** Consider a treatment effect model with an endogenous binary treatment D and with the potential outcome determined by a latent Roy model. The Roy model has also been studied extensively from both applied and theoretical perspectives. See for example the literature survey in Heckman and E.Vytlacil (2007) and the seminal paper in Heckman and Honoré (1990).

Here the observed outcome consists of two pieces: a continuous measure $Y = DY_1 + (1 - D)Y_0$ and a discrete indicator $W = DW_1 + (1 - D)W_0$ for $d = 0, 1$. These potential outcomes are given by

$$Y_d = \max_{j \in \{a,b\}} y^*_{j,d} \text{ and } W_d = \arg\max_{j \in \{a,b\}} y^*_{j,d}$$

where a and b index potential outcomes realized in different sectors, with

$$y^*_{j,d} = v_j(X, d) + \varepsilon_j.$$

The binary endogenous treatment D is determined as in the selection equation (1.2). For example, $D \in \{1, 0\}$ indicates whether an individual participates in certain professional training program, $W_d \in \{a, b\}$ indicates the potential sector in which the individual is employed, $y^*_{j,d}$ is the potential wage from sector j under treatment $D = d$, and $Y_d \in \mathbb{R}$ is the potential wage if the treatment status is $D = d$. As before, we maintain that $(X, Z) \perp (\varepsilon, U)$.

The parameter of interest is

$$\Pr\{Y_1 \leq y, W_1 = a | X\}$$

which by the independence $(X, Z) \perp (\varepsilon, U)$ and an application of the law of total probability can be decomposed into directly identifiable quantities and a counterfactual quantity

$$\begin{aligned} & \Pr\{Y_1 \leq y, W_1 = a \mid X = x, Z = z, D = 0\} \\ = \; & \Pr\left\{v_b(x, 1) + \varepsilon_b < v_a(x, 1) + \varepsilon_a \leq y \mid U \geq P(z)\right\}. \end{aligned} \tag{3.1}$$

Again, we seek to identify this counterfactual quantity by finding $\tilde{x} \in S_0$ such that

$$v_a(x, 1) = v_a(\tilde{x}, 0) \text{ and } v_b(x, 1) = v_b(\tilde{x}, 0) \tag{3.2}$$

This would allow us to recover the right hand side of (3.1) as

$$\Pr\{Y_0 \leq y, W_0 = a \mid X = \tilde{x}, Z = z, D = 0\}.$$

To find such a pair of $(x, \tilde{x})$, define $h_{d,W}(x,p,p'), h^*_{d,W}(x,p)$ by replacing $1\{Y \leq y\}$ with $1\{W = a\}$ in the definition of h_d, h^*_d in Section 2. Similarly, define $h_{d,Y}(x,y,p,p'), h^*_{d,Y}(x,y,p)$ by replacing $1\{Y \leq y\}$ with $1\{Y \leq y, W = a\}$ in the definition of h_d, h^*_d in Section 2. Then

$$\begin{aligned} h_{d,W}(x,p_1,p_2) &= \int_{p_2}^{p_1} \Pr\{v_b(x,d) + \varepsilon_b < v_a(x,d) + \varepsilon_a | U = u\} du; \\ h_{d,Y}(x,y,p_1,p_2) &= \int_{p_2}^{p_1} \Pr\{v_b(x,d) + \varepsilon_b < v_a(x,d) + \varepsilon_a \leq y | U = u\} du; \end{aligned}$$

and $h_{d,W}(x,p_1,p_2)$ and $h_{d,Y}(x,y,p_1,p_2)$ are both identified over their respective domains by construction. Assume $(\varepsilon_a, \varepsilon_b)$ is continuously distributed with positive density over $\mathbb{R}^2$ conditional on all u. Then the statement

$$\begin{aligned} \text{“}h_{1,W}(x,p,p') &= h_{0,W}(\tilde{x},p,p') \textit{ and } h_{1,Y}(x,y,p,p') = h_{0,Y}(\tilde{x},y,p,p') \\ \text{for all } y \text{ and } p &> p' \text{ on } P_x \cap P_{\tilde{x}}\text{”} \end{aligned}$$

holds true if and only if (3.2) holds. Then matching $h_{1,W}(x,p,p') = h_{0,W}(\tilde{x},p,p')$ ensures

$$v_a(x,1) - v_b(x,1) = v_a(\tilde{x},0) - v_b(\tilde{x},0); \tag{3.3}$$

while matching $h_{1,Y}(x,y,p,p') = h_{0,Y}(\tilde{x},y,p,p')$ *at the same time* ensures that in addition to (3.3)

$$v_a(x,1) = v_a(\tilde{x},0). \tag{3.4}$$

Combining (3.3) and (3.4) is equivalent to (3.2).

4 Extension

The identification strategy we have used so far requires matching exogenous variables $x, \tilde{x}$ on S_0, S_1. In some cases, with the outcome being continuous, we can construct similar argument for identifying a counterfactual quantity in a treatment effect model by matching different elements on the support of continuous outcome. This approach was not investigated in Vytlacil and Yildiz (2007), which focused on the use of first moment of outcome. The following example illustrates this point.

Example 4. (Potential outcome with random coefficients) Random coefficient models are prominent in both the theoretical and applied econometrics literature. They permit a flexible way to allow for conditional heteroscedasticity and unobserved heterogeneity. See, for example Hsiao and Pesaran (2008) for a survey. Here we consider a treatment effect model where the potential outcome is determined through random coefficients:

$$Y = DY_1 + (1 - D)Y_0 \text{ where } Y_d = (\alpha_d + X'\beta_d) \text{ for } d = 0, 1$$

and the binary endogenous treatment D is determined as in the selection equation (1.2). The *random* intercepts $\alpha_d \in \mathbb{R}$ and the *random* vectors of coefficients β_d are given by

$$\alpha_d = \bar{\alpha}_d(X) + \eta_d \text{ and } \beta_d = \bar{\beta}_d(X) + \varepsilon_d$$

where for any x and $d \in \{0, 1\}$., $(\bar{\alpha}_d(x), \bar{\beta}_d(x)) \in \mathbb{R}^{K+1}$ is a vector of constant parameters while $\eta_d \in \mathbb{R}$ and $\varepsilon_d \in \mathbb{R}^K$ are unobservable noises.

As in Vytlacil and Yildiz (2007), assume some elements in Z in the selection equation are excluded from X. We allow the vector of unobservable terms $(\epsilon_1, \epsilon_0, \eta_0, \eta_1, U)$ to be arbitrarily correlated. We also assume that

$$(X, Z) \perp (\epsilon_1, \epsilon_0, \eta_0, \eta_1, U), \tag{4.1}$$

with the marginal distribution of U normalized to standard uniform, so that $\theta(Z)$ is directly identified as $P(Z) \equiv E(D|Z = z)$.

In this example our goal is to identify the conditional distribution of Y_d given $X = x$ for $d = 0, 1$. From this result we can identify parameters of interest such as average treatment effects, quantile treatment effects, etc. Let $G_{P|x}$ denote the conditional distribution of $P \equiv P(Z)$ given $X = x$, which is directly identifiable from the data-generating process. By construction,

$$\Pr\{Y_1 \le y | X = x\} = \int \Pr\{Y_1 \le y | X = x, P = p\} dG_{P|x}(p),$$

where

$$\begin{aligned} & \Pr\{Y_1 \le y | X = x, P = p\} \\ = \; & E\left[D1\{Y_1 \le y\} | X = x, P = p\right] + E\left[(1 - D)1\{Y_1 \le y\} | X = x, P = p\right]. \end{aligned}$$

The first term on the right-hand side is identified as

$$E[D1\{Y \le y\} | X = x, P = p],$$

while the second term is counterfactual and can be written as

$$
\begin{aligned}
& \phi_0(x,y,p) \equiv E[1\{U \geq P\}1\{\alpha_1 + X'\beta_1 \leq y\}|X=x, P=p] \\
= \; & E[1\{U \geq p\}1\{\bar{\alpha}_1(x) + \eta_1 + x'(\bar{\beta}_1(x) + \varepsilon_1) \leq y\}] \\
= \; & \int_p^1 \Pr\{\eta_1 + x'\epsilon_1 \leq y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x)|U=u\}du.
\end{aligned}
$$

For any p on the support of P given $X = x$, define

$$
\begin{aligned}
& h_1^*(x,y,p) \equiv E\left[D1\left\{Y \leq y\right\}|X=x, P=p\right] \\
= \; & E\left[1\left\{U < P\right\}1\left\{\alpha_1 + X'\beta_1 \leq y\right\}|X=x, P=p\right] = E\left[1\left\{U < p\right\}1\left\{\alpha_1 + x'\beta_1 \leq y\right\}\right] \\
= \; & \int_0^p \Pr\{\eta_1 + x'\epsilon_1 \leq y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x)|U=u\}du,
\end{aligned}
$$

where the second equality uses (4.1). Likewise, under (4.1) we have:

$$
\begin{aligned}
& h_0^*(x,y,p) \equiv E\left[(1-D)1\left\{Y \leq y\right\}|X=x, P=p\right] \\
= \; & \int_p^1 \Pr\{\eta_0 + x'\epsilon_0 \leq y - \bar{\alpha}_0(x) - x'\bar{\beta}_0(x)|U_i=u\}du.
\end{aligned}
$$

Assume[4]

$$
F_{(\eta_1,\epsilon_1)|U=u} = F_{(\eta_0,\epsilon_0)|U=u} \text{ for all } u \in [0,1]. \tag{4.2}
$$

Under (4.2), we have

$$
\phi_0(x,y,p) = \int_p^1 \Pr\{\eta_0 + x'\epsilon_0 \leq y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x)|U=u\}du. \tag{4.3}
$$

Suppose for each pair (x, y) we can find $t(x, y)$ such that

$$
y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x) = t(x,y) - \bar{\alpha}_0(x) - x'\bar{\beta}_0(x).
$$

Then by construction

$$
\begin{aligned}
h_0^*(x,t(x,y),p) \; & \equiv \; \int_p^1 \Pr\{\eta_0 + x'\epsilon_0 \leq t(x,y) - \bar{\alpha}_0(x) - x'\bar{\beta}_0(x)|U=u\}du \\
& = \; \int_p^1 \Pr\{\eta_0 + x'\epsilon_0 \leq y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x)|U=u\}du = \phi_0(x,y,p)
\end{aligned}
$$

[4]This type of distributional equality assumption generalizes the exact equality of ϵ_1, ϵ_0 as can be found in for example Vytlacil and Yildiz (2007). Distributional equality has been used to motivate the *rank similarity* condition imposed frequently in the econometrics literature- see for example Chernozhukov and Hansen (2005), Frandsen and Lefgren (2018), Dong and Shen (2018), Chen and Khan (2014).

because of (4.3). Thus the counterfactual $\phi_0(x, y, p)$ would be identified as $h_0^*(x, t(x, y), p)$.

It remains to show that for each pair (x, y) we can uniquely recover $t(x, y)$ using quantities that are identifiable in the data-generating process. To do so, we define two auxiliary functions as follows: for $p_1 > p_2$ on the support of P given $X = x$, let

$$\begin{aligned} h_1(x, y, p_1, p_2) &\equiv h_1^*(x, y, p_1) - h_1^*(x, y, p_2) \\ &= \int_{p_2}^{p_1} \Pr\{\eta_1 + x'\epsilon_1 < y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x) | U = u\} du; \end{aligned}$$

and

$$\begin{aligned} h_0(x, y, p_1, p_2) &\equiv h_0^*(x, y, p_2) - h_0^*(x, y, p_1) \\ &= \int_{p_2}^{p_1} \Pr\{\eta_0 + x'\epsilon_0 < y - \bar{\alpha}_0(x) - x'\bar{\beta}_0(x) | U = u\} du. \end{aligned}$$

Suppose $\eta_d + x'\epsilon_d$ is continuously distributed over $\mathbb{R}$ for all values of x conditional on all $u \in [0, 1]$. Then for any fixed pair (x, y) and $p_1 < p_2$,

$$h_1(x, y, p_1, p_2) = h_0(x, t(x, y), p_1, p_2)$$

if and only if

$$t(x, y) = y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x) + \bar{\alpha}_0(x) + x'\bar{\beta}_0(x).$$

To see this, suppose $t(x, y) > y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x) + \bar{\alpha}_0(x) + x'\bar{\beta}_0(x)$, then (4.2) implies that $h_0(x, t(x, y), p_1, p_2) > h_1(x, y, p_1, p_2)$. A symmetric argument establishes a similar statement with "$>$" replaced by "$<$". This establishes our desired result.

5 Estimation

Here we outline estimation procedures from a random sample of the observed variables that are motivated by our identification results. We will first describe an estimation procedure for the parameter $E[Y_1]$ in the first three examples. Let P_x to denote the support of $P(Z)$ given $X = x$, $f_p(.|x)$ denote the density of $P(Z)$ given $X = x$, and

$$P_x^c = \{p\colon\ f_p(p|x) > c\} \text{ for a known } c > 0,$$

and for simplicity assume a strong overlap condition that

$$1 - c_0 > P(Z) > c_0 \text{ for a known } c_0 > 0,$$

Define a measure of distance between $h_1(x_1, \cdot)$ and $h_0(x_0, \cdot)$

$$\begin{aligned} &\|h_1(x_1, \cdot) - h_0(x_0, \cdot)\| \\ = \; &\left\{ \int\int\int \left(\int_{p_1}^{p_2} (F_{g|u}(y; v(x_1, 1)) - F_{g|u}(y; v(x_0, 0)) du \right)^2 I\left(p_1, p_2 \in P_x^c\right) w(y) dy dp_1 dp_2 \right\}^{1/2} \end{aligned}$$

where $w(y)$ is a chosen weight function. Consider the case when $h_1(x, y, p_1, p_2)$, $h_1(x, y, p_1, p_2)$ and $P(z)$ are known. For any given X_i, let $\tilde{X}_i$ be such that

$$\left\| h_1(X_i, \cdot) - h_0(\tilde{X}_i, \cdot) \right\| = 0$$

which, under Assumption A-4 in Section 2, is equivalent to

$$v(X_i, 1) = v(\tilde{X}_i, 0).$$

Define

$$\hat{Y}_i = E(Y|D = 0, \|h_1(X_i, \cdot) - h_0(X, \cdot)\| = 0, P = P_i).$$

Note that the conditional expectation on the right-hand side is equal to $E[Y|D = 0, v(X, 0) = v(X_i, 1), P = P_i]$, which in turn equals $E[Y_1|D = 0, X = X_i, P = P_i]$. Then, following the discussion above, we define the following estimator for $\Delta \equiv E[Y_1]$:

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^{n} \left(D_i Y_i + (1 - D_i) \hat{Y}_i \right)$$

or a **weighted version**

$$\hat{\Delta}_w = \frac{\frac{1}{n} \sum_{i=1}^{n} 1\left\{X_i \in A\right\} \left(D_i Y_i + (1 - D_i) \hat{Y}_i \right)}{\frac{1}{n} \sum_{i=1}^{n} 1\left\{X_i \in A\right\}}$$

Limiting distribution theory for each of these estimators follows from identical arguments in Vytlacil and Yildiz (2007). Here we formally state the theorem for the first estimator:

Theorem 5.1 *Under Assumptions A-1 to A-5, and the additional assumption that Y_1 has positive and finite second moment, then we have*

$$\sqrt{n}(\hat{\Delta} - \Delta) \xrightarrow{d} \mathbb{N}(0, V)$$

where

$$V = Var(E[Y_1|X, P, D]) + E[PVar(Y_1|X, P, D = 1)]$$

Now we describe an estimation procedure for the distributional treatment effect in Example 4, where we had a model with random coefficients. In this case, the parameter of interest is for a chosen value of the scalar y,

$$\Delta_2(y) = \Pr\{Y_1 \le y\}.$$

First, for fixed values of y and $p_1 > p_2$, we propose to estimate $t(x, y)$ as

$$\hat{t}(x, y, p_1, p_2) = \arg\min_t (h_1(x, y, p_1, p_2) - h_0(x, t, p_1, p_2))^2$$

and then average over values of p_1, p_2:

$$\hat{\tau}(x, y) = \frac{1}{n(n-1)} \sum_{i \neq j} I[P_i > P_j] \hat{t}(x, y, P_i, P_j)$$

An infeasible estimator for the parameter $\Delta_2(y)$, which assumes $t(x, y)$ is known, would be

$$\hat{\Delta}_2(y) = \frac{1}{n} \sum_{i=1}^{n} (D_i 1\{Y_i \le y\} + (1 - D_i) 1\{Y_i \le t(X_i, y)\}).$$

In practice, for feasible estimation, one needs to replace $t(x, y)$ by its estimator $\hat{\tau}(x, y)$.

6 Simulation Study

This section presents simulation evidence for the performance of the proposed estimation procedures described in Section 5, for both the Average Treatment Effect and the Distributional Treatment Effect. We report results for both our proposed estimator and that in Vytlacil and Yildiz (2007), for several designs. These include designs where the said monotonicity condition fails, and designs where the disturbance terms in the outcome equation are multidimensional.

Throughout all designs we model the treatment or dummy endogenous variable as

$$D = I[Z - U > 0]$$

where Z, U are independent standard normal. We experiment with the following designs for the outcome

Design 1

$$Y = X + 0.5 \cdot D + \epsilon$$

where X is standard normal, (ϵ, U) are distributed bivariate normal, each with mean 0 and variance 1, with correlations of 0,0.25,0.5.

Design 2

$$Y = X + 0.5 \cdot D + (X + D) \cdot \epsilon$$

where X is distributed standard normal, (ϵ, U) are distributed bivariate normal, each with mean 0 and variance 1, with correlations of 0,0.25,0.5.

Design 3

$$Y = (X + 0.5 \cdot D + \epsilon)^2$$

where X is distributed standard normal, (ϵ, U) are distributed bivariate normal, each with mean 0 and variance 1, with correlations of 0,0.25,0.5.

We note that the monotonicity condition is satisfied in design 1 but fails in the other two designs. For each of these designs, we report results for estimating the parameter $E[Y_1]$, which denotes the expected value for potential outcome under treatment $D = 1$. The two estimators used in the simulation study were the one proposed in Section 5 and the method proposed in Vytlacil and Yildiz (2007). The summary statistics, scaled by the true parameter value, Mean Bias, Median Bias, Root Mean Squared Error, (RMSE), and Median Absolute Deviation (MAD) were evaluated for sample sizes of 100, 200, 400 for 401 replications. Results for each of these designs are reported in Tables 1 to 3 respectively. In implementing our procedure we assumed the propensity score function is known, and conducted next stage estimation using a nonparametric kernel estimator with normal kernel function, and a bandwidth of $n^{-1/5}$. This rate reflects "undersmoothing" as there are two regressors, the propensity score and the regressor X. For the estimator in Vytlacil and Yildiz (2007), which involved the derivative of conditional expectation functions as well, estimating these functions nonparametrically gave very unstable results so we report results for an infeasible version of their estimator, assuming such functions, as well as the propensity scores, are known.

To implement the second stage of our proposed procedure, in calculating the distance $\|h_1(x_i, \cdot) - h_0(x_0, \cdot)\|$ we used an evenly space grid of values for y, and selected $n/50$ grid points, with n denoting the sample size.

The results indicate the desirable properties of our proposed procedure, generally agreeing with Theorem 5.1. In all designs our estimator has small values for bias and RMSE, with the value of RMSE decreasing as the sample size grows. In contrast, the procedure based on Vytlacil and Yildiz (2007) only performs well in Design 1, with values of bias and RMSE comparable to those using our method. As in our procedure these values decrease with as the sample size grows, which is expected, as the monotonicity condition rely on is satisfied in these designs. In this case, their approach has smaller standard errors largely due to the relative simpler structure of the infeasible version, but their biases persist even when the sample size increases.

For designs 2 and 3, where monotonicity is violated, the procedure proposed in Vytlacil and Yildiz (2007) does not perform well. In design 2 in Table 2 both the bias and RMSE are generally increasing with the sample size. Results for their estimator are better in design 3, but the bias hardly converges with the sample size and is much larger compared to our estimator.

We also simulate data from a model with dummy endogenous variable and potential outcomes determined by random coefficients. It is important to note that for this design, the original matching idea in Vytlacil and Yildiz (2007) does not apply. This is because different values of x lead to different distribution of the composite error $\eta_d + x'\epsilon_d$. Our contribution in Section 4 is to propose a new approach based on matching different values of outcome y, rather than the regressors x. Based on the counterfactual framework discussed in Section 4, here the treatment variable D is modeled as the same way as the dummy endogenous variable above. Similarly the regressor X is standard normal. For both Y_0, Y_1 the random intercepts were modeled as constants (0 and 1, respectively) and the additive error terms were each standard normal. For the random slopes, the means were 1 and 2 respectively, and the additive error terms were also standard normal, independent of all other disturbance terms and each other. Here we use the procedure in Section 4 to estimate the parameter $\Delta_2 = P(Y_1 < y)$, where in the simulation we set $y = 1$. The same four summary statistics are reported for sample sizes 100,200,400, based on 401 replications. Results for this random coefficients design are reported in Table 4.

The estimator proposed in Section 5 performs well; but the bias and RMSE are much small at 400 observations compared to 100 and 200 observations, indicating convergence at the parametric rate.

Table 1

	CKT			VY		
ρ_v	0	1/4	1/2	0	1/4	1/2
n=100						
MEAN BIAS	-0.0170	0.0229	-0.0435	-0.1302	-0.1676	-0.2018
MEDIAN BIAS	-0.0137	0.0124	-0.0653	-0.1318	-0.1678	-0.2087
RMSE	0.4936	0.4800	0.4945	0.3308	0.3337	0.3546
MAD	0.3289	0.3328	0.3156	0.2200	0.2271	0.2546
n=200						
MEAN BIAS	0.0032	-0.0024	-0.0069	-0.0864	-0.1299	-0.1766
MEDIAN BIAS	-0.0102	-0.0141	-0.0314	-0.0934	-0.1277	-0.1679
RMSE	0.3355	0.3367	0.3521	0.2293	0.2457	0.2711
MAD	0.2240	0.2228	0.2517	0.1594	0.1676	0.1865
n=400						
MEAN BIAS	-0.0187	0.0101	-0.0055	-0.0584	-0.11134	-0.1593
MEDIAN BIAS	-0.0261	0.0128	-0.0065	-0.0592	-0.1162	-0.1572
RMSE	0.2496	0.2489	0.2578	0.2049	0.1867	0.2167
MAD	0.1523	0.1732	0.1659	0.1197	0.1345	0.1605

Table 2

	CKT			VY		
ρ_v	0	1/4	1/2	0	1/4	1/2
n=100						
MEAN BIAS	0.0109	0.0397	-0.0671	-0.1509	-0.2875	-0.4207
MEDIAN BIAS	0.0151	0.0227	-0.0939	-0.1590	-0.2918	-0.4262
RMSE	0.5089	0.2737	0.4853	0.3524	0.4199	0.5289
MAD	0.3395	0.2447	0.3105	0.2419	0.30898	0.4310
n=200						
MEAN BIAS	0.0322	0.0143	-0.0311	-0.1273	-0.2559	-0.3875
MEDIAN BIAS	0.0159	0.0054	-0.0543	-0.1310	-0.2553	-0.3884
RMSE	0.3487	0.3444	0.3455	0.2622	0.3407	0.4475
MAD	0.2317	0.2297	0.2552	0.1782	0.2624	0.3884
n=400						
MEAN BIAS	0.0088	0.0269	-0.0294	-0.0962	-0.2247	-0.3708
MEDIAN BIAS	0.0007	0.0244	-0.0309	-0.0982	-0.2255	-0.3769
RMSE	0.2578	0.2557	0.2549	0.1920	0.2764	0.4037
MAD	0.1649	0.1733	0.1606	0.1354	0.2283	0.3769

Table 3

	CKT			VY		
ρ_v	0	1/4	1/2	0	1/4	1/2
n=100						
MEAN BIAS	-0.0097	-0.0070	0.0019	-0.0691	-0.0898	-0.1066
MEDIAN BIAS	-0.0233	-0.0101	-0.0240	-0.0799	-0.0925	-0.1178
RMSE	0.1893	0.2085	0.2126	0.1546	0.1630	0.1701
MAD	0.1398	0.1342	0.1374	0.1125	0.1178	0.1315
n=200						
MEAN BIAS	-0.0108	-0.0069	-0.0068	-0.0609	-0.0765	-0.0968
MEDIAN BIAS	-0.0148	-0.0033	-0.0099	-0.0674	-0.0769	-0.1017
RMSE	0.1372	0.1434	0.1424	0.1163	0.1262	0.1369
MAD	0.0949	0.0989	0.0953	0.0855	0.0887	0.1078
n=400						
MEAN BIAS	-0.0073	-0.0014	-0.0026	-0.0583	-0.0725	-0.0889
MEDIAN BIAS	-0.0149	-0.0023	-0.0029	-0.0610	-0.0751	-0.0887
RMSE	0.1084	0.0994	0.0989	0.0924	0.1007	0.1131
MAD	0.0697	0.0685	0.0654	0.0689	0.0788	0.0901

Table 4

CKT			
ρ_v	0	1/4	1/2
n=100			
MEAN BIAS	0.0109	-0.0086	0.0038
MEDIAN BIAS	0.0000	-0.0064	0.0126
RMSE	0.1011	0.0979	0.0955
MAD	0.0600	0.0648	0.0652
n=200			
MEAN BIAS	-0.0050	-0.0150	0.0095
MEDIAN BIAS	-0.0100	-0.0161	0.0029
RMSE	0.0669	0.0669	0.0665
MAD	0.0400	0.0454	0.0457
n=400			
MEAN BIAS	0.0012	-0.0132	0.0074
MEDIAN BIAS	0.0049	-0.0162	0.0077
RMSE	0.0501	0.0494	0.0495
MAD	0.0349	0.0325	0.0360

7 Conclusion

In this paper, we considered identification and estimation of nonseparable models with endogenous binary treatment. Existing approaches are based on a monotonicity condition, which is violated in models with multiple unobserved idiosyncratic shocks. Such models arise in many important empirical settings, including Roy models and multinomial choice models with dummy endogenous variables, as well as treatment effect models with random coefficients. We establish novel identification results for these models which are constructive and conducive to estimation procedures which are easy to compute and whose limiting distributional properties follow from standard large sample theorems. A simulation study indicates adequate finite sample performance of our proposed methods.

This paper leaves open areas for future research. Our method requires the selection of the number and location of cutoff points, so a data driven method for selecting these would be useful. Furthermore, the relative efficiency of our proposed approach needs to be explored, perhaps by deriving efficiency bounds for these new classes of models.

References

Ahn, H., and J. Powell (1993): "Semiparametric Estimation of Censored Selection Models," *Journal of Econometrics*, 58, 3–29.

Ahn, H., J. Powell, H. Ichimura, and P. Ruud (2017): "Simple Estimators for Invertible Index Models," *Journal of Business Economics and Statistics*, 36, 1–10.

AUERBACH, E. (2019): "Identification and Estimation of a Partially Linear Regression Model using Network Data," mimeograph, Northwestern University.

CHEN, S., S. KHAN, AND X. TANG (2016): "On the Informational Content of Special Regressors in Heteroskedastic Binary Response Models," *Journal of Econometrics*, 193, 162–182.

CHEN, S. H., AND S. KHAN (2014): "Semiparametric Estimation of Program Impacts on Dispersion of Potential Wages," *Journal of Applied Econometrics*, 29, 901–919.

CHERNOZHUKOV, V., AND C. HANSEN (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245 – 261.

D'HAULTFOEUILLE, X., AND P. FEVRIER (2015): "Identification of Nonseparable Triangular Models with Discrete Instruments," *Econometrica*, 3, 1199–1210.

DONG, Y., AND S. SHEN (2018): "The Empirical Content of the Roy Model," *The Review of Economics and Statistics*, 100, 78–85.

FENG, J. (2020): "Matching Points: Supplementing Instruments with Covariates in Triangular Models," mimeograph, Columbia University.

FRANDSEN, B., AND L. LEFGREN (2018): "Testing Rank Similarity," *The Review of Economics and Statistics*, 100, 86–91.

HAN, S., AND E. J. VYTLACIL (2017): "Identification in a generalization of bivariate probit models with endogenous regressors," *Journal of Econometrics*, 199, 63–73.

HECKMAN, J., AND E.VYTLACIL (2007): "Econometric evaluation of social programs," in *Handbook of Econometrics, Vol. 6B*, ed. by J. Heckman, and E. Leamer. Amesterdam: North Holland.

HECKMAN, J., AND B. HONORÉ (1990): "The Empirical Content of the Roy Model," *Econometrica*, 58, 1121–1149.

HSIAO, C., AND M. H. PESARAN (2008): "Random Coefficient Models," in *The Econometrics of Panel Data: Advanced Studies in Theoretical and Applied Econonmetrics*, ed. by L. Matyas, and P. Sevestre. Springer.

IMBENS, G., AND W. NEWEY (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77(5), 1481–1512.

Khan, S., A. Maurel, and Y. Zhang (2019): "Informational Content of Factor Structures in Simultaneous Discrete Response Models," Working Paper, Duke University.

Khan, S., F. Ouyang, and E. Tamer (2019): "Inference in Semiparametric Mutinomial Response Models," Boston College Working Paper.

Lee, L.-F. (1995): "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics*, 65, 381–428.

Lewbel, A., D. Jacho-Chavez, and J. Encarnciono (2016): "Identification and estimation of semiparametric two-step models," *Quantitative Economics*, 7, 561–589.

Mourifié, I. (2015): "Sharp Bounds on Treatment Effects in a Binary Triangular System," *Journal of Econometrics*, 187(1), 74–81.

Pakes, A., and J. Porter (2014): "Moment Inequalties for Multinomial Choice with Fixed Effects," Harvard University Working Paper.

Ruud, P. (2000): "Semiparametric estimation of discrete choice models," mimeograph, University of California at Berkeley.

Shaikh, A. M., and E. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations with Binary Dependent Variables," *Econometrica*, 79(3), 949–955.

Shi, X., M. Shum, and W. Song (2018): "Estimating Semi-Parametric Panel Multinomial Choice Models using Cyclic Monotonicity," *Econometrica*, 86, 737–761.

Torgovitsky, A. (2015): "Identification of Nonseparable Models Using Instruments with Small Support," *Econometrica*, 3, 1185–1197.

Vuong, Q., and H. Xu (2017): "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity," *Quantitative Economics*, pp. 589–610.

Vytlacil, E. J., and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75, 757–779.