Identification and Identification Failure for Treatment Effects using Structural Systems

Halbert White*  Karim Chalak
UC San Diego  Boston College

December 29, 2011

Abstract We provide necessary and sufficient conditions for effect identification, thereby characterizing the limits to identification. Our results link the non-structural potential outcome framework for identifying and estimating treatment effects to structural approaches in economics. This permits economic theory to be built into treatment effect methods. We elucidate the sources and consequences of identification failure by examining the biases arising when the necessary conditions fail, and we clarify the relations between unconfoundedness, conditional exogeneity, and the necessary and sufficient identification conditions. A new quantity, the exogeneity score, plays a central role in this analysis, permitting an omitted variable representation for effect biases. This analysis also provides practical guidance for selecting covariates and insight into the price paid for making various identifying assumptions and the benefits gained.

JEL Classification Numbers: C10, C20, C30, C50.

Keywords: conditional exogeneity; covariate selection; exogeneity score; nonparametric estimation; simultaneous equations; structural equations; treatment effects; unconfoundedness.

This is a preprint of an article that has been accepted for publication in Econometric Reviews.

1 Introduction

One of the major concerns in econometrics is to identify and estimate the effects of one set of endogenous variables on another. The pioneering work of the Cowles Commission laid foundations, mostly cast in terms of linear systems, that have shaped econometric theory and practice ever since. There, exclusion restrictions and exogenous instruments, correlated with the causes of interest, but uncorrelated with the system's unobservable disturbances, play the central role in identifying and estimating the effects of interest.

*Corresponding author. Address: Department of Economics 0508, UCSD, La Jolla, CA 92093-0508. Phone: 858 534-3502; Fax 858 534-7040; Email: hwhite@ucsd.edu. This is a revised version of a paper previously titled "Identifying Structural Effects in Nonseparable Systems Using Covariates."
Over the years, econometricians have worked to relax the conditions under which such effects can be identified and estimated. One strand of literature analyzes nonlinear or nonparametric separable systems.\textsuperscript{1} As separability rules out interactions between observable and unobservable drivers of the response of interest, another strand of the literature analyzes nonparametric nonseparable systems, typically obeying certain monotonicity restrictions.\textsuperscript{2}

The instrumental variables approach runs into difficulties, however, in fully general nonparametric nonseparable structures. As Hahn and Ridder (2011) show for the method of instrumental regression (see Darolles, Fan, Florens, and Renault, 2011), and as Schennach, White, and Chalak (2011) show for derivative ratio methods (generalizations of Haavelmo’s (1943) indirect least squares (ILS)), separability in the structural equations governing the endogenous causes of interest plays a crucial role in the identification of average marginal effects when using exogenous instruments. IV estimation methods generally do not work in fully nonseparable structures.

If economists are to draw reliable policy inferences and gain useful economic insight from data without placing possibly erroneous separability restrictions on the structures thought to generate the data, other methods, not using exogenous instrumental variables, must be employed. One powerful class of such methods employs conditioning variables referred to variously as "covariates" or "control variables" to achieve effect identification. In particular, in the treatment effects/potential outcome literature (e.g., Rubin, 1974; Rosenbaum and Rubin, 1983), the key identifying condition of \textit{unconfoundedness} requires the potential outcomes to be independent of treatment, given the covariates.

These methods have been developed and applied in economics to great effect, especially in labor economics.\textsuperscript{3} In sharp contrast to the treatment effect literature, a significant strand of the economics literature specifies explicit structure involving unobservables. In this context, the identifying condition is that the causes of interest (corresponding to treatments) are independent of the unobservable drivers, given specific control variables (covariates). Parallel to the classical Cowles Commission approach, this conditional independence is interpreted as a form of exogeneity;

\textsuperscript{1}For example, Brown (1983) studies parametric separable systems nonlinear only in regressors. Newey and Powell (2003), Darolles, Fan, Florens, and Renault (2011), and Santos (2011), among others, study nonparametric separable systems.

\textsuperscript{2}Chesher (2003, 2005), Imbens and Newey (2009), Matzkin (2003, 2008), Chernozhukov and Hansen (2005), Chernozhukov, Imbens, and Newey (2007), and Hoderlein, Su, and White (2011), for example, consider nonparametric nonseparable systems that impose certain monotonicity conditions.

for convenience, we refer to this as conditional exogeneity. Recent examples of this approach are the studies by Altonji and Matzkin (2005), Hoderlein and Mammen (2007), and Hoderlein (2011). Also, a significant development is the use of covariates to identify and estimate not only average effects of causes of interest (treatment) but also their effects on various aspects of the response distribution, such as quantiles (e.g., Firpo, 2007).

With this background, the main contributions of this paper can be clearly articulated. First, we directly link the non-structural potential outcome/treatment effect approach to the structural equations approach traditionally used by economists. This permits economic theory to be built into treatment effect methods. Second, we characterize the limits to effect identification in either approach by providing necessary and sufficient conditions for identification. Third, we elucidate the sources and consequences of identification failure by examining the biases arising when the necessary conditions fail. A new quantity, the exogeneity score, plays a central role in this analysis, permitting an omitted variable representation for effect biases. Finally, we clarify the relations between unconfoundedness, conditional exogeneity, and the necessary and sufficient identification conditions. Further useful consequences of these investigations are: (1) practical guidance for selecting covariates, which can otherwise be a murky process; (2) an appreciation of the price paid for making various identifying assumptions and the benefits gained by paying this price; and (3) insight into recovering direct and indirect effects of treatment, as well as total effects.

The plan of the paper is as follows. In Section 2 we briefly review the treatment effect approach, highlighting the use of Imbens’s (2000) weak unconfoundedness condition to identify standard treatment effects. We also call attention to the difficulty there in justifying covariates and to the challenges to examining the sources and consequences of identification failure. Section 3 introduces a structural system compatible with both the treatment effect approach and standard econometric structural systems, including systems satisfying simultaneous equations. This makes clear the relation between the latter and the treatment effect approach; it also suggests some interesting new estimators. We show that once the triangularity implicit in the treatment effect approach is made explicit, the structural system of Section 3 is formally equivalent to the data generating process underlying the treatment effect approach.

Section 4 studies identification and identification failure using the structural system of Section 3. We explicitly consider not only average effects of binary treatments, but also average marginal effects of continuous treatments. We give necessary and sufficient conditions for identification, showing that neither weak unconfoundedness nor conditional exogeneity is necessary for identifi-
cation. We also show the central role played by the exogeneity score in determining effect biases and governing possibilities for near identification of effects when identification fails. Section 5 provides guidance for the selection of covariates, based on the previous sections.

Section 6 extends the results of Section 4 to cover not just average effects of treatment, but also effects of treatment on general aspects of the response distribution. Section 7 discusses how these general distributional effects can be estimated via quasi-maximum likelihood or generalized method of moments (GMM) methods, which may be parametric, nonparametric, or semiparametric. Section 8 examines the relations between weak unconfoundedness and conditional exogeneity, and Section 9 summarizes and concludes.

As we will be blending the treatment effect and structural equation frameworks, each of which has its own familiar notation, we face a non-trivial notational challenge. We have attempted to chart a middle course in selecting notation, with the hoped-for result that those with preference for one notation or another will be equally (dis)satisfied.

2 Identification of Treatment Effects

We first briefly review identification in the treatment effect context. For further details, we refer the reader to the valuable surveys of Imbens (2004) and Imbens and Wooldridge (2009). We also draw on Imbens (2000) and Hirano and Imbens (2004). We then discuss identification failure.

We begin by specifying the data generating process (DGP).

Assumption A.1 Let covariates X and treatments D be finitely dimensioned random vectors with supports $\mathcal{X}$ and $\mathcal{D}$. For each $d$ in $\mathcal{D}$, let $Y_d$ be a random scalar, the potential outcome at $d$.

For a given observation from a population governed by $A_1$, we do not observe the entire collection of potential outcomes $\{Y_d, \, d \in \mathcal{D}\}$. Instead, we only observe the outcome $Y_d$ for the actual treatment, $d$. For example, with binary treatment $D$ (i.e., $\mathcal{D} = \{0, 1\}$), we observe $Y = Y_0 1\{D = 0\} + Y_1 1\{D = 1\}$, where $1\{\cdot\}$ is the indicator function, equal to 1 if the event in brackets is true and to 0 otherwise.

---

4All random vectors are defined on an underlying complete probability space $(\Omega, \mathcal{F}, P)$ that we will mostly leave implicit. Completeness means that all subsets of sets with $P$–measure zero are measurable. The support of a random variable is the smallest closed set containing the random variable with probability 1. For example, $\mathcal{D} := \text{supp}(D)$ is such that $P[D \in \mathcal{D}] = 1$. The notation $D$ can be considered a mnemonic for "dose".
Interest attaches to a variety of treatment effects. In the binary case, leading examples are

\[
\begin{align*}
\beta &:= E(Y_1 - Y_0) & \text{Average Effect of Treatment} \\
\gamma_0 &:= E(Y_1 - Y_0 \mid D = 0) & \text{Average Effect of Treatment on the Untreated} \\
\gamma_1 &:= E(Y_1 - Y_0 \mid D = 1) & \text{Average Effect of Treatment on the Treated}
\end{align*}
\]

Because we observe only the outcome for the actual treatment and not other potential outcomes, these effects cannot be recovered without further information. One possibility is to use the \textit{weak unconfoundedness} condition (Imbens, 2000; Hirano and Imbens, 2004):

**Assumption A.2** \( Y_d \perp D \mid X \) for all \( d \in \mathcal{D} \).

We follow Dawid (1979), writing \( X \perp Y \mid Z \) when \( X \) is independent of \( Y \) given \( Z \). Assumption A.2 weakens Rosenbaum and Rubin’s (1984) unconfoundedness condition, \( \{Y_d, d \in \mathcal{D}\} \perp D \mid X \).

In experimental contexts, A.2 is guaranteed when treatment is random, conditional on \( X \).

Consider recovering \( \gamma_1 \). We have

\[
\gamma_1 = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1)
\]

\[
= \{E(Y \mid D = 1) - \mu_{0,1}\} + \{\mu_{0,1} - E(Y_0 \mid D = 1)\}, \quad \text{with}
\]

\[
\mu_{0,1} := E[ E(Y \mid D = 0, X) \mid D = 1].
\]

Given A.1 and A.2, we have \( \mu_{0,1} = E(Y_0 \mid D = 1) \), as we verify shortly. If this holds, then

\[
\gamma_1 = E(Y \mid D = 1) - \mu_{0,1}.
\]

Following Hurwicz (1950), we say \( \gamma_1 \) is \textit{identified}, as it can be represented entirely in terms of the joint distribution of the observables, here \( D, Y, \) and \( X \).

To verify \( \mu_{0,1} = E(Y_0 \mid D = 1) \) under A.2, apply the law of iterated expectations to write

\[
E(Y_0 \mid D = 1) = E[ E(Y_0 \mid D = 1, X) \mid D = 1] = E[ E(Y_0 \mid X) \mid D = 1]
\]

\[
= E[ E(Y_0 \mid D = 0, X) \mid D = 1]
\]

\[
= E[ E(Y \mid D = 0, X) \mid D = 1] =: \mu_{0,1}.
\]

The second and third equalities follow from A.2 and the fourth from the definition of \( Y \).

For easy reference, we summarize familiar identification results for the binary treatment case:

**Proposition 2.1** Suppose A.1 and A.2 hold and \( E(Y) < \infty \). If \( D \) is a binary scalar, then \( \beta, \gamma_0, \) and \( \gamma_1 \) are identified as
\[
\begin{align*}
\beta &= \gamma_1 \ P[D = 1] + \gamma_0 \ P[D = 0] \\
\gamma_0 &= \mu_{1,0} - E(Y \mid D = 0) \quad \text{and} \quad \gamma_1 = E(Y \mid D = 1) - \mu_{0,1} \quad \text{with} \\
\mu_{1,0} &= E[ E(Y \mid D = 1, X) \mid D = 0].
\end{align*}
\]

Such identification results immediately suggest ways to estimate the effects of interest from a sample, via what Efron has called the plug-in principle: replace unknown population quantities by their various sample counterparts. Once identification is established, the rest is "just" statistics.

As noted above, Assumption A.2 can be guaranteed in controlled experiments by random treatment, conditional on \(X\). In non-experimental contexts, it may similarly be thought that A.2 holds for some given choice of covariates \(X\). We call such situations "quasi-experimental." See Imbens (2004, section VI) and Imbens and Wooldridge (2009, p. 23-24) for several examples.

Significantly, however, the justification for specifying a particular set of covariates in quasi-experimental settings is opaque, as conditional random assignment of treatment is either not possible or, when it is possible, non-compliance with treatment assignment is also a strong possibility. Nor is there an explicit mechanism within the treatment effect approach for specifying how the \(X\)'s arise. This calls A.2 into question, so it is important to determine what it is that "apparent" treatment effects (the quantities on the right in Proposition 2.1) recover when Assumption A.2 fails.

To address this issue, again consider \(\gamma_1\). We write the apparent effect as

\[
E(Y \mid D = 1) - \mu_{0,1} = \{E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1)\} + \{E(Y_0 \mid D = 1) - \mu_{0,1}\} \\
= \gamma_1 + b_1,
\]

where \(b_1 := E(Y_0 \mid D = 1) - \mu_{0,1}\) is the associated effect bias. When A.2 holds, \(b_1 = 0\), so the apparent effect coincides with the true effect, \(\gamma_1\). When A.2 fails, however, we generally have identification failure: \(b_1 \neq 0\). To investigate the source of this failure, we write

\[
\begin{align*}
b_1 &= E(Y_0 \mid D = 1) - E[E(Y \mid D = 0, X) \mid D = 1] \\
&= E[E(Y_0 \mid D = 1, X) \mid D = 1] - E[E(Y_0 \mid D = 0, X) \mid D = 1] \\
&= E(b_1^*(X) \mid D = 1), \quad \text{where} \\
b_1^*(X) &= E(Y_0 \mid D = 1, X) - E(Y_0 \mid D = 0, X).
\end{align*}
\]

Clearly, \(b_1^*(x)\) measures the degree of departure from unconfoundedness for potential outcome \(Y_0\) at \(X = x\), so \(b_1\) measures the average degree of this departure for the treated. But interpreting
\( \beta_1(X) \) raises a significant challenge, as this requires insight into \( E(Y_0 \mid D = 1, X) \), the expected potential outcome in the absence of treatment, given \( X \) and the presence of treatment. Without unconfoundedness, this is likely not an intuitively informative object for most researchers.

To gain better insight into identification failure, we seek additional sources of information that can provide deeper understanding. For economists, economic theory is a natural source of such information. Thus, we examine the use of economic structure to provide insight into necessary and sufficient conditions for treatment effect identification.

### 3 Structural Foundations for Treatment Effects

We specify a triangular structural system compatible with the treatment effect framework that also encompasses standard econometric systems of structural equations.

**Assumption B.1** Let a triangular system generate finitely dimensioned observable random vectors \( W \) and \( Z \) and countably dimensioned random vectors \( U \) and \( V \), and suppose that finitely dimensioned observable vectors \( Y \) and \( D \) are determined as

\[
Y = r(D, Z, U) \quad D = q(Z, W, V),
\]

where the "structural functions" \( q \) and \( r \) are unknown measurable functions.

This notation corresponds to that of the previous section: \( Y \) is the observable outcome, \( D \) is the observable treatment or "cause of interest"; the other observables are \( X := (W, Z) \). We take \( Y \) and \( D \) to have positive dimension. The other variables may have dimension zero. Conventionally, \( U \) and \( V \) are designated as unobservable, but it is better to view these as variables that will be omitted from empirical analysis. They may be omitted because they are unobservable, but it may also be that the researcher has purposefully or inadvertently neglected them. We also allow some latitude in designating \( Z \) and \( W \) as observable. It may be that these require sample estimation and are observable only in the large-sample limit; this will cause no difficulty here.

Explicit discussion of cause and effect may be uncomfortable for some readers. In triangular structural systems, however, there is no ambiguity in equating the statement "\( D \) causes \( Y \) " with the statement "\( Y \) is a non-constant structural function of \( D \)." White and Chalak (2009) and Chalak and White (2011b) give rigorous foundations. The presence of \( D, Z, \) and \( U \) in \( r(\cdot) \) means that these are potential, not mandatory, causes of \( Y \), and similarly for \( Z, W, \) and \( V \) appearing in
q. Although not explicit in B.1, we adopt the convention that D does not cause U; if it did, we could perform substitutions that would yield a structure satisfying B.1, but with different U’s.

How do B.1 and A.1 relate? By A.1, \( Y_d \) is a random variable for each \( d \). Thus, its realizations can always be written \( Y_d(\tilde{u}) \), where \( \tilde{u} \) is the realization of some random variable, say \( \tilde{U} \). Defining \( r(d, \tilde{u}) := Y_d(\tilde{u}) \), A.1 implies \( Y = r(D, \tilde{U}) \). For given \( \tilde{u} \), \( r(\cdot, \tilde{u}) \) is the dose-response function. We observe only \( y = r(d, \tilde{u}) \), the realized outcome for the treatment actually applied, as the treatment effect setup requires. Similarly, A.1 specifies that D is a random vector, so we can always write \( D = q(\tilde{V}) \), for some random variable \( \tilde{V} \) and some measurable function \( q \).

What about \( X \)? A.1 specifies only that \( X \) is a random vector. B.1 also specifies this, but it seems to go further, by specifying how elements of \( X \) determine \( Y_d \) and \( D \). One aspect of B.1 merely assigns labels: If some elements of \( X \) can potentially affect both \( Y_d \) and \( D \), then we label them \( Z \). If other elements of \( X \) cannot affect \( Y_d \), we label them \( W \). As the \( Z \)’s and \( W \)’s are not mandatory causes, their appearance in \( r \) and \( q \) is not a restriction. Writing \( \tilde{U} := (Z, U) \) and \( \tilde{V} := (Z, W, V) \) then gives

\[
Y = r(D, \tilde{U}) = r(D, Z, U) \quad D = q(\tilde{V}) = q(Z, W, V).
\]

B.1 is more explicit than A.1, however, in specifying triangularity. But it is always understood that there is a directional link from treatment to response, and not the reverse. This ensures triangularity from \( D \) to \( Y \). Further, whatever the other drivers of \( Y \) may be, they must precede \( Y \) in time, justifying triangularity from \( \tilde{U} := (Z, U) \) to \( Y \). It is also understood that neither the treatment (Rosenbaum, 1984) nor the response should drive the covariates; this ensures triangularity with respect to \( X := (Z, W) \) and \((Y, D)\). Triangularity from \( V \) to \( D \) is ensured by the fact that unobserved treatment drivers \( V \) precede \( D \) in time. Thus, B.1 makes explicit the triangularity assumptions implicit in the treatment effect approach. With these understood, A.1 implies B.1.\(^5\)

Conversely, B.1 implies A.1: Given B.1, \( X \) is a random vector, and measurability of \( q \) ensures \( D \) is a random vector. For each \( d \in D \), let \( Y_d := r(d, Z, U) \). Measurability of \( r \) ensures \( Y_d \) is a random variable, so B.1 implies A.1. Further, B.1 enforces the triangularity implicit in the treatment effect approach. With triangularity understood, A.1 is equivalent to B.1: the treatment effect approach is formally equivalent to the structural approach.

Assumption B.1 accommodates the structures of economic theory. A leading example is the generalized Roy model (e.g., Heckman and Vytlacil, 2005; Heckman, 2010). Often, economic

\(^5\)B.1 also requires \( r \) to be jointly measurable. Without this, \( Y \) can fail to be a random variable, rendering probability calculations and expectations meaningless. This technical requirement is thus also implicit in A.1.
theory suggests variables besides $D$ that may drive $Y$. In B.1, this is accommodated by the presence of $Z$ and $U$. If these drivers can be faithfully observed, we may assign them to $Z$. Otherwise, we assign them to $U$. Economics also typically specifies exclusion restrictions – that certain variables do not determine $Y$. Here, $W$ and $V$ are excluded variables, as they do not appear inside $r(\cdot)$. Nevertheless, B.1 does not demand exclusion restrictions: $W$ and $V$ may have dimension zero.

B.1 also makes explicit the structure generating $D$, as is common in econometric structural systems (e.g., Vytlacil, 2002). Triangularity is enforced, as $Y$ does not appear in $q$. Representing the structure generating $D$ is also useful in the treatment effect context, as this can formalize both treatment assignment and possible non-compliance (e.g., Imbens and Angrist, 1994; Angrist, Imbens, and Rubin, 1996). For example, if treatment is randomly assigned conditional on observables $X := (Z, W)$, then $D = q(Z, W, V)$ can represent this assignment, where $V$ codes the randomizing factor. To handle non-compliance, we can write

$$D = q(Z, W, V) = q_1(q_0(Z, W, V_0), Z, W, V_1),$$

where $D_0 := q_0(Z, W, V_0)$ is the randomly assigned treatment, $q_1(D_0, Z, W, V_1)$ maps the assigned treatment to the applied treatment, $D$, and $V := (V_0, V_1)'$.

To see that B.1 accommodates a broad range of specific economic structures encountered in the literature, first consider the linear triangular system

$$Y = B_{21}D + \Gamma_{21}Z + U \quad D = \Gamma_{11}Z + \Gamma_{12}W + V,$$

where $\Gamma_{11}$, $\Gamma_{12}$, $\Gamma_{21}$, and $B_{21}$ are unknown coefficient matrices. This is a classical recursive structure of the form intensively studied by the Cowles Commission and their followers. Clearly,

$$r(D, Z, U) = B_{21}D + \Gamma_{21}Z + U \quad q(Z, W, V) = \Gamma_{11}Z + \Gamma_{12}W + V.$$

A more general structure is the separable system

$$Y = r_0(D, Z) + U \quad D = q_0(Z, W) + V,$$

where $q_0$ and $r_0$ need not be linear. An extension of the separable system has the form of B.1, but assumes that $q(z, w, \cdot)$ and/or $r(d, z, \cdot)$ are strictly monotone in a scalar unobservable for each $(d, z, w)$. (See footnote 2 above for relevant citations.)

The system specified by B.1 permits each of these structures, but does not impose linearity, separability, or monotonicity in a scalar unobservable. This is important, because although economic theory can readily specify which variables matter, which don’t, and where, it is often much
less specific about such details of functional form. This flexibility also reflects the philosophy of the treatment effect approach, which avoids imposing unwarranted structure.

B.1 encompasses not just triangular systems but also nonrecursive systems. In particular, a major strand of econometrics is concerned with simultaneous equations. B.1 is fully compatible with such systems. Specifically, suppose \( Y = (Y_1', Y_2')' \) satisfies the simultaneous equations\(^6\)

\[
Y_1 = s_1(Y_2, D, Z, U) \quad Y_2 = s_2(Y_1, D, Z, U).
\]

When a fixed point for this system is unique or can be suitably selected, it represents "full equilibrium" and has the form \( Y = r(D, Z, U) \). We write this more explicitly as

\[
Y_1 = r_1(D, Z, U) \quad Y_2 = r_2(D, Z, U).
\]

The simultaneous system ensures that

\[
r_1(D, Z, U) = s_1(r_2(D, Z, U), D, Z, U)
\]

\[
r_2(D, Z, U) = s_2(r_1(D, Z, U), D, Z, U).
\]

Under suitable assumptions, the full equilibrium effects of \( D \) on \( Y_1 \) and \( Y_2 \) can be used to recover "partial equilibrium" effects, for example, the marginal effect of \( Y_2 \) on \( Y_1 \) embedded in \( s_1 \).

To see this, suppose that \( Y_1 \) and \( Y_2 \) are \( \ell_1 \times 1 \) and \( \ell_2 \times 1 \), respectively, and that \( D \) is \( k \times 1 \), with a \( k_2 \times 1 \) subvector \( D_2 \) that appears only in \( s_2 \) and not in \( s_1 \) (the analog of the classical instrument exclusion restriction). Differentiating the first equation above with respect to \( d_2 \) gives

\[
\nabla_{d_2} r_1(d, z, u)_{k_2 \times \ell_1} = \nabla_{d_2} r_2(d, z, u)_{k_2 \times \ell_2} \nabla_{y_2} s_1(r_2(d, z, u), d, z, u)_{\ell_2 \times \ell_1},
\]

where \( \nabla_{d_2} \) and \( \nabla_{y_2} \) are the gradient operators with respect to \( d_2 \) and \( y_2 \), respectively. Next, pre-multiply both sides of this equation by \( \nabla_{d_2} r_2(d, z, u)' \) to get

\[
\nabla_{d_2} r_2(d, z, u)' \nabla_{d_2} r_1(d, z, u) = [\nabla_{d_2} r_2(d, z, u)' \nabla_{d_2} r_2(d, z, u)] \nabla_{y_2} s_1(r_2(d, z, u), d, z, u).
\]

Provided \( \nabla_{d_2} r_2(d, z, u)' \nabla_{d_2} r_2(d, z, u) \) has full rank (this is the generalized rank condition, for which the order condition \( k_2 \geq \ell_2 \) is necessary), we can solve this as

\[
\nabla_{y_2} s_1(r_2(d, z, u), d, z, u) = [\nabla_{d_2} r_2(d, z, u)' \nabla_{d_2} r_2(d, z, u)]^{-1} \nabla_{d_2} r_2(d, z, u)' \nabla_{d_2} r_1(d, z, u).
\]

\(^6\)Although \( s_1 \) and \( s_2 \) embody causal economic structure, say for partial equilibrium, we follow Strotz and Wold (1960) by not viewing simultaneous equations as causal relations; this obviates instantaneous causality. Instead, we interpret these as mutual consistency conditions, e.g., characterizing equilibria, thereby linking partial and full equilibrium. White and Chalak (2009) give further discussion.
These are the partial equilibrium marginal effects of $Y_2$ on $Y_1$ at the equilibrium value $y_2 = r_2(d, z, u)$. Identification of $\nabla y_2 r_1$ and $\nabla y_2 r_2$ implies identification of $\nabla y_2 s_1$ and suggests some novel estimators. Our investigation below of the identification of $\nabla d r$ is directly relevant here.

This discussion provides a new link between the treatment effect approach and classical systems of simultaneous equations. On the one hand, it suggests that treatment effect-based methods may be productively applied to systems of nonparametric simultaneous equations (see e.g. Angrist, Graddy, and Imbens, 2000). On the other, it suggests that when multiple outcomes from treatment are jointly determined, it may be possible to infer the partial equilibrium effect of one of the outcomes on the other, as if one of the outcomes were subject to intervention. We return to this in Section 7, where we discuss estimation.

4 Effect Identification in Triangular Structural Systems

Weak unconfoundedness identifies effects of interest in the treatment effect context. We now examine identification and identification failure for the structural system of B.1.

The first result verifies that random treatment given $X$ ensures unconfoundedness, implying weak unconfoundedness. To implement random treatment given $X$, we choose $V$ so that $V \perp U \mid X$. It suffices for this that $V \perp (U, X)$, as can be easily arranged in controlled experiments.

**Proposition 4.1** Given B.1, if $V \perp U \mid X$, then $D \perp U \mid X$. This implies unconfoundedness, \{\{Y_d, d \in D\} \perp D \mid X, which in turn implies weak unconfoundedness, $Y_d \perp D \mid X$ for all $d \in D$.

This demonstrates a concrete benefit of the structural approach: the genesis of the covariates $X := (W, Z)$ is clarified. We now see that the covariates include other drivers $Z$ of $Y$, as well as other observables $W$ that may drive $D$, but do not drive $Y$. By triangularity, $W$ is not driven by $Y$ or $D$ but may be driven by $Z, U, or V$. Interestingly, when some elements of $W$ do not cause $D$ but are instead driven by $U$, then these covariates can be observed after treatment or even after the response. We further discuss covariate choice in Section 5 below.

We can have $D \perp U \mid X$ without random treatment, so this is often taken as primitive:

**Assumption B.2** $D \perp U \mid X$.

Examples are Altonji and Matzkin (2005), Hoderlein and Mammen (2007), Imbens and Newey (2009), and Hoderlein, Su, and White (2011). Altonji and Matzkin (2005) assume the covariates
X do not enter r, so Z is null; Hoderlein and Mammen (2007) assume that all covariates enter r, so W is null. Assumption B.2 allows both possibilities. When X has dimension zero, this is classical strict exogeneity of D, D ⊥ U, so we refer to B.2 as "conditional" exogeneity of D. Because of the instrumental role of X in ensuring identification, we follow Chalak and White (2011a) in calling X "conditioning instruments."

4.1 Marginal Effects

Propositions 4.1 and 2.1 imply that when B.1 and B.2 hold with binary D, then β, γ₀, and γ₁ are identified. When D is continuously distributed and r is suitably differentiable, interest attaches to a variety of average marginal effects, as in Altonji and Matzkin (2005), Hoderlein and Mammen (2007), and Hoderlein (2011). The structural approach provides a convenient context for analyzing such effects. When D is scalar, consider the average marginal effect on Y of D at d given X = x,

\[ \tilde{\gamma}_d(x) := E(\nabla_d r(D, Z, U) \mid D = d, X = x) = E(\lim_{\epsilon \to 0} \frac{Y_{d + \epsilon} - Y_d}{\epsilon} \mid D = d, X = x). \]

The final expression represents this effect in terms of potential outcomes, showing that this is a covariate-conditioned marginal analog of γ₀ and γ₁. Given (D, X) = (d, x), \( \tilde{\gamma}_d(x) \) is the mean squared error-optimal predictor of \( \nabla_d r(D, Z, U) \), the marginal effect of D. In the separable case, \( r(D, Z, U) = r_0(D, Z) + U \), \( \tilde{\gamma}_d(x) \) is the marginal effect itself, as \( \tilde{\gamma}_d(x) = \nabla_d r_0(d, z) = \nabla_d r(d, z, u) \).

The direct analog of γ₀ and γ₁ is \( \bar{\beta} := E(\bar{\gamma}_d(X)) \). The average derivative of Stoker (1986) and Powell, Stock, and Stoker (1989) is then \( \bar{\beta} := E(\bar{\gamma}_D) \), the analog of β.

In the separable case, B.1 and B.2 easily suffice for identification, as

\[ \nabla_d E(Y \mid D = d, X = x) = \nabla_d r_0(d, z) + \nabla_d E(U \mid X = x) = \nabla_d r_0(d, z). \]

The nonseparable case is more involved. Under B.1, B.2, and some regularity conditions, we have

\[ \tilde{\gamma}_d(x) = E(\nabla_d r(D, Z, U) \mid D = d, X = x) = E(\nabla_d r(d, Z, U) \mid X = x) = \nabla_d E(r(d, Z, U) \mid X = x) \]

\[ = \nabla_d E(Y \mid D = d, X = x). \]

The second and fourth equalities hold by B.2 and the third equality holds under an interchange of expectation and derivative.
So far, the identification of $\tilde{\gamma}_d(x)$ is purely heuristic. To definitively establish identification for $\tilde{\gamma}_d(x)$ requires explicit conditions justifying the interchange of derivative and integral. Imbens and Newey (2009, theorem 6) give such conditions for the case where $Z$ is absent, ensuring the desired identification. Here, we state closely related conditions that serve as a prototype below.

We now take $D$ to be scalar-valued for simplicity. If $D$ is vector-valued, we can interpret $Z$ in B.3 as including elements of $D$ other than that whose marginal effect is of interest.

**Assumption B.3** Given $d \in \text{int}(D)$ and $x := (w, z) \in \mathcal{X}$, let $\mathcal{C}$ be a non-empty open subset of $D \subseteq \mathbb{R}$ containing $d$. (i) $E(r(d, Z, U) \mid X = x) < \infty$; (ii) For all $\epsilon$ such that $d + \epsilon \in \mathcal{C}$, $\nabla \epsilon r(d + \epsilon, z, u)$ exists for almost every\(^7\) $u$; (iii) There is a random variable $\Delta_z(U)$ with $E(\Delta_z(U) \mid X = x) < \infty$ such that for all $\epsilon$ such that $d + \epsilon \in \mathcal{C}$, $|\nabla \epsilon r(d + \epsilon, z, u)| \leq \Delta_z(u)$ for almost every $u$.

Requiring that $\mathcal{C}$ is non-empty ensures that $D$ is continuously distributed on $\mathcal{C}$. B.3 is fairly straightforward to state for the structural approach, but less so in the treatment effect context. A complement to Imbens and Newey (2009, theorem 6) is

**Theorem 4.2** Suppose B.1 and B.2 hold.

(i) if B.3 holds for given $d$ and $x$, then

$$\tilde{\gamma}_d(x) = \nabla_d E(Y \mid D = d, X = x);$$

(ii) if B.3 holds for given $d$ and all $x$ in $\mathcal{X}$, then

$$\tilde{\gamma}_d = E[\nabla_d E(Y \mid D = d, X)];$$

(iii) if B.3 holds for all $d$ in $\text{int}(D)$ and all $x$ in $\mathcal{X}$, then

$$\bar{\beta} = E[\nabla_d E(Y \mid D, X)].$$

Part (i) is a local identification result at $(d, x)$. Parts (ii) and (iii) make use of local identification for all $x$ and for all $(d, x)$, respectively, to verify identification of less specific effect measures.

### 4.2 Identification Failure and Near Identification

What happens if B.2 fails? We first consider marginal effects; then we return to the binary case.

\(^7\)The qualifier "almost every" (a.e.) here means that the condition can be violated for $u$ belonging to a measurable set $\mathcal{V}$ having $P[U \in \mathcal{V} \mid X = x] = 0$. 

13
4.2.1 Marginal Effects

For marginal effects, local identification failure of $\tilde{\gamma}_d(x)$ on a set of positive $(D, X)$—probability generally results in identification failure for $\tilde{\gamma}_d$ and $\tilde{\beta}$. Thus, we focus on local identification.

We begin by considering the separable case, where without B.2 we have

$$\nabla_d E(Y \mid D = d, X = x) = \nabla_d r_0(d, z) + \nabla_d E(U \mid D = d, X = x).$$

The effect bias is

$$\tilde{\beta}_d^* (x) := \nabla_d E(U \mid D = d, X = x).$$

Thus, $\tilde{\beta}_d^* (x) = 0$ is the necessary and sufficient condition for local identification with separability.

Clearly, B.2 ensures $\tilde{\beta}_d^* (x) = 0$ for all $(d, x)$ but is not necessary. Here, a weaker sufficient exogeneity condition is $E(U \mid D, X) = E(U \mid X)$, conditional mean independence of $U$ from $D$, given $X$. Indeed, this ensures global identification of $\tilde{\gamma}_d(x)$, in the sense that $\tilde{\gamma}_d(x)$ is locally identified for all $(d, x)$. Conditional mean independence is not necessary for local identification, as $\tilde{\beta}_d^* (x) = 0$ can hold even when $P[ E(U \mid D, X) = E(U \mid X) ] < 1$.

Stronger structural restrictions do permit weaker exogeneity conditions. In the linear case,

$$Y = B_{21} D + \Gamma_{21} Z + U \quad \text{so} \quad \nabla_d r(d, z, u) = \nabla_d r_0(d, z) = B_{21}'.$$

In this case, the effect of interest, $B_{21}$, can be recovered by linear projection (see Chalak and White, 2011a), as

$$B_{21}^{**} := [E([D - E(D \mid X)] D')]^{-1} E([D - E(D \mid X)] Y')$$

$$= B_{21} + [E([D - E(D \mid X)] D')]^{-1} E([D - E(D \mid X)] U')$$

$$= B_{21}' + \tilde{\beta},$$

so the necessary and sufficient condition for identification in the linear case is $\tilde{\beta} = 0$. That is, identification holds if and only if $U$ is uncorrelated with the regression errors $D - E(D \mid X)$.

Here, B.2 is sufficient for this, as is $E(U \mid D, X) = E(U \mid X)$; but neither is necessary.

We see a clear trade-off: the stronger the structural restrictions, the weaker the exogeneity conditions ensuring identification. Thus, in the general nonseparable case, we expect stronger exogeneity conditions. For this case, let $G(u \mid d, x) := P[U \leq u \mid D = d, X = x]$ define the

\footnote{For convenience, we call this "conditional non-correlation". Note that $E([D - E(D \mid X)] U') = 0$ is equivalent to $E([D - E(D \mid X)] [U' - E(U' \mid X)]) = 0$.}
conditional CDF of \( U \) given \((D, X) = (d, x)\), and let \( \mathcal{U}_{d,x} \) be the support of \( U \) given \((D, X) = (d, x)\).

Throughout, we assume that conditional distributions are regular (Dudley, 2002, ch. 10.2), so

\[
E(Y \mid D = d, X = x) = \int_{\mathcal{U}_{d,x}} r(d, z, u) \, dG(u \mid d, x).
\]

Differentiating this with respect to \( d \) yields two sets of terms from the integral: one contributed by the integrand and another by the boundary of \( \mathcal{U}_{d,x} \). For now, we simplify by taking \( \mathcal{U}_{d,x} = \mathcal{U}_x \) for all \( d \), removing the complications associated with the boundary terms.

Next, we replace \( dG(\cdot \mid d, x) \) with \( g(\cdot \mid d, x) \, d\nu(\cdot \mid x) \), where \( g(\cdot \mid d, x) \) is the Radon-Nikodym density (see, e.g., Bartle, 1966, theorem 8.9) and \( \nu(\cdot \mid x) \) is a \( \sigma \)-finite measure not depending on \( d \) (for now), dominating the measure defined by \( G(B \mid d, x) = \int_B dG(u \mid d, x) \). This gives

\[
E(Y \mid D = d, X = x) = \int_{\mathcal{U}_x} r(d, z, u) \, g(u \mid d, x) \, d\nu(u \mid x).
\]

Differentiating, interchanging integral and derivative, and applying the product rule give

\[
\nabla_d E(Y \mid D = d, X = x) = \int \nabla_d r(d, z, u) \, g(u \mid d, x) \, d\nu(u \mid x) + \int r(d, z, u) \, \nabla_d g(u \mid d, x) \, d\nu(u \mid x)
\]

\[
= \tilde{\gamma}_d(x) + \int r(d, z, u) \, \nabla_d \ln g(u \mid d, x) \, g(u \mid d, x) \, d\nu(u \mid x)
\]

\[
= \tilde{\gamma}_d(x) + \tilde{b}_d(x),
\]

where we suppress \( \mathcal{U}_x \), use the fact that \( \nabla_d g = (\nabla_d \ln g) \, g \), and write the local effect bias as

\[
\tilde{b}_d(x) = \int r(d, z, u) \, \nabla_d \ln g(u \mid d, x) \, g(u \mid d, x) \, d\nu(u \mid x).
\]

Thus, the necessary and sufficient exogeneity condition for local identification is \( \tilde{b}_d(x) = 0 \).

Our next result sharpens this heuristic analysis. Because we seek necessary and sufficient conditions for identification, attention to technical detail becomes crucial. Assumption B.4 of the Appendix imposes a milder domination condition on \( G(\cdot \mid d, x) \) than above, ensuring that for all \( \delta \) in a neighborhood of \( d \), \( \mathcal{U}_{d,x} = \mathcal{U}_{d,x} \) and that \( G(u \mid \delta, x) = g(u \mid \delta, x) \, d\nu(u \mid d, x) \) for all \( u \in \mathcal{U}_{d,x} \). Assumption B.5 modifies B.3 to accommodate the absence of B.2. Assumption B.6 of the Appendix justifies defining

\[
\tilde{b}_d(x) := \int_{\mathcal{U}_{d,x}} r(d, z, u) \, \nabla_d \ln g(u \mid d, x) \, dG(u \mid d, x).
\]

---

\(^9\) Domination means that whenever \( \nu(B \mid x) = 0 \), then \( G(B \mid d, x) = 0 \), written \( G(\cdot \mid d, x) \ll \nu(\cdot \mid x) \). If \( U \) is continuous, its distribution is dominated by Lebesgue measure, and \( g \) is a continuous probability density function. If \( U \) is discrete, its distribution is dominated by counting measure, and \( g \) is a probability mass function.
Theorem 4.3 Suppose that B.1 holds with $E(Y) < \infty$, and that B.4 - B.6 hold for $d$ and $x$ as specified below. Then

(i) for given $x \in \mathcal{X}$ and $d \in \text{int}(D)$, $\tilde{b}_d(x) = 0$ if and only if

$$\tilde{\gamma}_d(x) = \nabla_d E(Y \mid D = d, X = x);$$

(ii) for given $d \in \text{int}(D)$, (a) if $\tilde{b}_d(x) = 0$ holds for all $x$ in $\mathcal{X}$, then

$$\tilde{\gamma}_d = E(\nabla_d E(Y \mid D = d, X));$$

(b) if $\tilde{\gamma}_d = E(\nabla_d E(Y \mid D = d, X))$, then with $\tilde{\beta}_d := E(\nabla_d E(Y \mid D = d, X)) - \tilde{\gamma}_d$,

$$\tilde{\beta}_d = E[1\{X \in \mathcal{X}_d\} \tilde{b}_d(X) \mid D = d] = 0,$$

where $\mathcal{X}_d := \{x \in \mathcal{X} : \tilde{b}_d(x) \neq 0\}.$

This result implies that although B.2 is sufficient for identification in the nonseparable case, it is by no means necessary. Instead, the crucial identification condition $\tilde{b}_d(x) = 0$ is a conditional non-correlation requirement, as we see in eq. (1) below. It can hold for a single $(d, x)$, ensuring local identification, as in (i); it can hold for given $d$ and all $x$, ensuring treatment-specific identification, as in (ii); or it can hold for all $(d, x)$, ensuring local identification and treatment-specific identification globally. B.2 is sufficient, but not necessary, for global identification.

Because $\tilde{b}_d(x) = 0$ is both necessary and sufficient, it is the weakest possible condition for local marginal effect identification for unrestricted structures. Part (ii) shows that $\tilde{b}_d(X) = 0 \ a.s.$ is essentially necessary for treatment-specific identification of $\tilde{\gamma}_d$, in the sense that if it fails, then identification requires some fortunate cancellations of local effect bias. We view such cancellations as a less viable foundation than $\tilde{b}_d(X) = 0 \ a.s.$ for claiming identification, so the latter is the weakest plausible condition ensuring identification of $\tilde{\gamma}_d$. For brevity, we do not state a result for $\tilde{\beta}$, but the situation is analogous: the weakest plausible condition ensuring identification of $\tilde{\beta}$ is $\tilde{\beta}_D(X) = 0 \ a.s.$

If $\tilde{b}_d(x) = 0$ suffices to identify $\tilde{\gamma}_d(x)$, why assume B.2? First, not only is B.2 straightforward and powerful, but it can be structurally justified in a variety of ways. Random treatment is just one possibility. Chalak and White (2011a,b) and White and Lu (2011a) provide many other examples. Second, and equally important, is that the condition $\tilde{b}_d(x) = 0$ relies on a possibly fortuitous orthogonality between $Y$ and a specific feature of the conditional distribution of the
unobservables, \( S := \nabla_d \ln g(U \mid D, X) \):

\[
\tilde{b}_d(x) = E(Y \mid D = d, X = x) = E(\varepsilon \mid D = d, X = x) = 0.
\]

It is often easier to justify B.2 than to justify this orthogonality a priori. Thus, although B.2 is not necessary, it is a convenient, powerful, and often plausible assumption for ensuring identification of marginal effects, and, as we will see, other effects as well.

Now consider what happens when even \( \tilde{b}_d(x) = 0 \) fails. The random variable \( S \) plays a key role, so we examine it closely. If B.2 holds, then \( S = 0 \) a.s. The converse also holds, so variations in \( S \) indicate departures from conditional exogeneity. We also have that for all \((d, x)\),

\[
E[S \mid D = d, X = x] = \int \nabla_d \ln g(u \mid d, x) \, g(u \mid d, x) \, dv(u \mid d, x) = \nabla_d g(u \mid d, x) \, dv(u \mid d, x) = 0.
\]

The last equality holds as \( \int g(u \mid d, x) \, dv(u \mid d, x) = 1 \) for all \((d, x)\). We add Assumption B.7 of the Appendix to help ensure that the interchange of derivative and integral holds. Thus, conditional variation in \( S \) signals local violations of conditional exogeneity. We call \( S \) the \textit{exogeneity score}, similar to the score in maximum likelihood estimation. As we will see, this quantity and its analogs play a central role in analyzing identification and its failure.

A key result follows from the fact that \( E[S \mid D = d, X = x] = 0 \), specifically that

\[
\tilde{b}_d(x) = E(\varepsilon S \mid D = d, X = x), \tag{1}
\]

where \( \varepsilon \) is the regression residual, \( \varepsilon := Y - E(Y \mid D, X) \). Thus, \( \tilde{b}_d(x) \) is a \textit{generalization of the traditional omitted variable bias, with omitted variable} \( S \).

Given sufficient moments, it follows by, for example, the Cauchy-Schwarz inequality that

\[
|\tilde{b}_d(x)| \leq \sigma_\varepsilon(d, x) \, \sigma_S(d, x),
\]

where \( \sigma_\varepsilon^2(d, x) := \text{var}(\varepsilon \mid D = d, X = x) \) and \( \sigma_S^2(d, x) := \text{var}(S \mid D = d, X = x) \). This is a local \textit{near} identification result: the effect bias becomes smaller, other things equal, as either the conditional regression variance or the local departure from conditional exogeneity decreases.

This inequality is sharp, as it can hold as an equality. To give an example, let \( r \) be separable with \( U \) conditionally normal, \( U \mid (D, X) \sim N(\mu(D, X), \sigma^2(D, X)) \). We saw above that \( \tilde{b}_d(x) = \nabla_d \mu(d, x) \). By separability, \( \varepsilon = U - \mu(D, X) \) and \( \sigma_\varepsilon^2(D, X) = \sigma^2(D, X) \). By normality,

\[
S = \nabla_d \mu(D, X) \frac{(U - \mu(D, X))}{\sigma_\varepsilon^2(D, X)} + \nabla_d \sigma_\varepsilon(D, X) \frac{(U - \mu(D, X))^2}{\sigma_\varepsilon^2(D, X)}, \quad \text{so}
\]

17
\[ \sigma_S^2(d, x) = \frac{\nabla_d \mu(d, x)^2 + 3 \nabla_d \sigma_\varepsilon(d, x)^2}{\sigma_\varepsilon^2(d, x)}. \]

The effect bias inequality gives

\[ |\tilde{b}_d(x)| = |\nabla_d \mu(d, x)| \leq \sigma_\varepsilon(d, x) \times (\nabla_d \mu(d, x)^2 + 3 \nabla_d \sigma_\varepsilon(d, x)^2)^{1/2}/\sigma_\varepsilon(d, x) \]

\[ = (\nabla_d \mu(d, x)^2 + 3 \nabla_d \sigma_\varepsilon(d, x)^2)^{1/2}. \]

This becomes an equality whenever \( \nabla_d \sigma_\varepsilon(d, x) = 0 \), which cannot be ruled out.

We have the following sharp near identification result.

**Corollary 4.4** Suppose that B.1 holds and that B.4 - B.7 hold for given \( d \) and \( x \). If \( \sigma_\varepsilon(d, x) \) and \( \sigma_S(d, x) \) are finite, then we have the sharp bound \( |\tilde{b}_d(x)| \leq \sigma_\varepsilon(d, x) \sigma_S(d, x) \).

The sharpness of the bound means that in the absence of conditions guaranteeing \( \tilde{b}_d(x) = 0 \), there are always structures satisfying B.1 for which identification fails. But the inequality also says that if the local departure from conditional exogeneity is small, so is the local effect bias. Alternatively, if the regression of \( Y \) on \((D, X)\) provides a very good fit at \((d, x)\), then the effect bias is again small. There is a trade-off: the better the local regression fit, the more of a local departure from conditional exogeneity can be tolerated to achieve a given bias magnitude.

### 4.2.2 Binary Treatment Effects

Now let \( D \) be binary, and denote the covariate-conditioned effect of treatment on the treated

\[ \gamma_1^*(X) := E(Y_1 \mid D = 1, X) - E(Y_0 \mid D = 1, X), \]

so that \( \gamma_1 := E(Y_1 - Y_0 \mid D = 1) = E(\gamma_1^*(X) \mid D = 1) \). Then

\[ E(Y \mid D = 1, X) - E(Y \mid D = 0, X) = \gamma_1^*(X) + b_1^*(X). \]

Recall that \( b_1^*(X) := E(Y_0 \mid D = 1, X) - E(Y_0 \mid D = 0, X) \) is the effect bias we sought to interpret in the absence of unconfoundedness in Section 2.

For binary \( D \), A.2 is stronger than necessary for identification. The next result gives necessary and sufficient exogeneity conditions for \( \gamma_1 \) and \( \gamma_1^*(x) \). Results for \( \gamma_0 \) and \( \gamma_0^*(x) \) are analogous.

**Proposition 4.5** Suppose that A.1 holds with \( E(Y) < \infty \) and \( D \) binary. Then

(i) For given \( x \in \mathcal{X} \), \( b_1^*(x) = 0 \) if and only if

\[ \gamma_1^*(x) = E(Y \mid D = 1, X = x) - E(Y \mid D = 0, X = x); \]
(ii) (a) If \( b_1^*(x) = 0 \) for all \( x \) in \( X \), then with \( \mu_{0,1} := E(E(Y \mid D = 0, X) \mid D = 1) \),

\[
\gamma_1 = E(Y \mid D = 1) - \mu_{0,1};
\]

(b) If \( \gamma_1 = E(Y \mid D = 1) - \mu_{0,1} \), then with \( X_1 := \{x \in X : b_1^*(x) \neq 0\} \),

\[
b_1 = E[ 1\{X \in X_1\} \ b_1^*(X) \mid D = 1 ] = 0.
\]

Note that \( b_1^*(x) = 0 \) for all \( x \) in \( X \) is equivalent to\(^{10} \) \( E[Y_0 \mid D, X] = E[Y_0 \mid X] \).

Proposition 4.5 shows that for the binary case, A.2 can be relaxed in three ways, and identification can still hold. To identify \( \gamma_1 \), we see from \((ii)\) that conditional independence of \( Y_0 \) from \( D \) given \( X \) can be relaxed to conditional mean independence (see also Heckman and Vytlacil, 2005, p. 708). Second, this relaxation need not hold for all \( d \), but just for specific \( d \) : in \((ii)\), \( \gamma_1 \) is identified without requiring \( E[Y_1 \mid D, X] = E[Y_1 \mid X] \). An analogous result holds for \( \gamma_0 \). Only if we wish to identify \( \beta \) do we generally need conditional mean independence for both \( d = 0 \) and \( d = 1 \). Third, we see that even when \( E[Y_0 \mid D, X] \neq E[Y_0 \mid X] \), we can still achieve identification of \( \gamma_1^*(x) \) for those \( x \) values where \( E[Y_0 \mid D, X = x] = E[Y_0 \mid X = x] \) does hold.

Examing necessity, \((i)\) shows that \( E[Y_0 \mid D, X = x] = E[Y_0 \mid X = x] \) cannot be further relaxed, or local identification fails. Part \((ii)\) shows that \( E[Y_0 \mid D, X] = E[Y_0 \mid X] \) (i.e., \( b_1^*(X) = 0 \) \( a.s. \)) is essentially necessary to identify \( \gamma_1 \), in the sense that if it fails, then identification requires some fortunate cancellations of local effect bias. Parallel to the marginal case, \( E[Y_0 \mid D, X] = E[Y_0 \mid X] \) is the weakest plausible condition ensuring identification of \( \gamma_1 \).

If \( E[Y_0 \mid D, X] = E[Y_0 \mid X] \) suffices to identify \( \gamma_1 \), why assume A.2? First, not only is A.2 convenient, but it weakens traditional unconfoundedness. Second, like conditional exogeneity, A.2 globally identifies a wide array of general effects, as we show later. A third motivation emerges from a structural representation for \( b_1^*(x) \), based on the binary exogeneity score, \( S_{1,0} := dG(U \mid 1, X)/dG(U \mid 0, X) - 1 : \)

\[
b_1^*(x) = E[Y_0 \mid D = 1, X = x] - E[Y_0 \mid D = 0, X = x] = \int r(0, z, u) \left[ \frac{dG(u \mid 1, x)}{dG(u \mid 0, x)} - 1 \right] dG(u \mid 0, x)
\]

\[
= E(Y S_{1,0} \mid D = 0, X = x).
\]

The integral expression above and the final expectation are structural representations that provide insight into the effect bias not as easily accessible from the treatment effect perspective. This

---

\(^{10}\)The intuitive content of this representation hinges on unconfoundedness. When \( b_1^*(X) \neq 0 \), we are still dealing with \( E(Y_0 \mid D = 1, X) - E(Y_0 \mid D = 0, X) \).
shows that the necessary condition $b_1^*(x) = 0$ relies on a possibly fortuitous orthogonality between $Y$ and the binary exogeneity score. It can be hard to justify this a priori. Thus, although A.2 is not necessary, it is a convenient, powerful, and often plausible general assumption.

In the representation above, $S_{1,0}$ plays the same role as that played by $S$ for marginal effects. We have $E(S_{1,0}) = 0$ and $E(S_{1,0} | X = x) = 0$ for all $x$. As for $S$, variation in $S_{1,0}$ indicates departures from conditional exogeneity (B.2), and conditional variation in $S_{1,0}$ indicates local departures. Letting $\varepsilon_0 := Y - E(Y | D = 0, X)$, we also have

$$b_1^*(X) = E(\varepsilon_0 S_{1,0} | D = 0, X),$$

the omitted variable bias representation. Cauchy-Schwarz then gives

$$|b_1^*(x)| \leq \sigma_{\varepsilon_0}(x) \sigma_{S_{1,0}}(x),$$

where $\sigma_{\varepsilon_0}^2(x) := \text{var}(\varepsilon_0 | D = 0, X = x)$ and $\sigma_{S_{1,0}}^2(x) := \text{var}(S_{1,0} | D = 0, X = x)$. This gives a local near identification result analogous to the marginal effect case, showing that the local treatment effect bias decreases with either the residual regression variance for the untreated case or the local degree of departure from conditional exogeneity. Again, the inequality is sharp.

Similar results hold for $\gamma_0$ and $\beta$. These are now straightforward, so we omit them.

5 Covariate Choice

Above, we noted that the covariates should include drivers $Z$ of $Y$ other than $D$; they may also include other observables $W$ that may drive $D$, that do not drive $Y$, that are not driven by $Y$ or $D$, and that may be driven by $Z$, $U$, or $V$. Although this provides essential guidance, it does not ensure the absence of effect bias. Further structure, such as random treatment, is required. The structural approach facilitates investigating this.

First, suppose $W$ is empty. Then B.2 becomes $D \perp U | Z$, as in Hoderlein and Mammen (2007). A simple sufficient condition for this is $(D, Z) \perp U$. This is the classical strict exogeneity assumption for $(D, Z)$. With separability, this can be weakened to $E(U | D, Z) = 0$. With linearity, this can be weakened to $U$ uncorrelated with $(D, Z)$. Arguably, most applied econometric work adopts one of these assumptions, either explicitly or implicitly, with or without justification.

To be succinct in this section, we mainly focus on the general nonseparable case and B.2. Nevertheless, our discussion can usually be translated to the separable or linear cases, substituting conditional mean independence or conditional non-correlation for conditional independence.
When \((D, Z) \perp U\), the effects of both \(D\) and \(Z\) on \(Y\) are identified. Indeed, we can absorb \(Z\) into \(D\) in this case, dispensing with covariates entirely. But when only the effects of \(D\) are of interest, the weaker condition \(D \perp U \mid Z\) still suffices. Random treatment conditional on observed drivers \(Z\) of \(Y\) is an example in which \((D, Z) \perp U\) can easily fail, but \(D \perp U \mid Z\) holds.

When \(D \perp U \mid Z\) fails, the possibility of using additional covariates \(W\) creates opportunities for identification. This has been exploited to great effect in structural econometrics and labor economics, including the work of Barnow, Cain, and Goldberger (1980), Heckman and Robb (1985), Heckman, Ichimura, and Todd (1998), and Heckman and Vytlacil (2005). Chalak and White (2011a, section 4.1.1) discuss four different structures ensuring that \(D \perp U \mid X\), with \(W\) non-empty. White and Lu (2011a) discuss four related structures ensuring \(B.2\), based on conditions considered by Hahn (2004):

\[(D, W) \perp U \mid Z \quad \text{and} \quad D \perp (U, W) \mid Z,
\]
both of which imply \(D \perp U \mid (Z, W)\). A further possibility is \((D, W_1) \perp (U, W_2) \mid Z\), as this also implies \(D \perp U \mid (Z, W)\), with \(W := (W_1, W_2)\). Chalak and White (2011b) discuss in depth the various ways that triangular structures can yield \(D \perp U \mid X\). For brevity, we do not summarize these results here; the interested reader can find details and further references in these articles.

One particularly interesting structure ensuring \(D \perp U \mid X\) is the triangular system

\[Y = r(D, U) \quad D = q(W, V),\]
with \(V \perp (U, W)\). This is the structure treated by Imbens and Newey (2009), with notation adapted to match Assumption B.1. Here, we interpret \(V\) as classical strictly exogenous instruments and \(W\) as drivers of \(D\) that are not directly observable. This is a nonparametric generalization of classical triangular structural systems. Now \(V \perp (U, W)\) implies \(V \perp U \mid W\). This implies \((V, W) \perp U \mid W\), which then implies \(D \perp U \mid W\). Thus, if \(W\) could be observed, we could take \(X = W\). Imbens and Newey (2009) use the fact that if \(W\) is a scalar and \(q(\cdot, v)\) is strictly monotonic for each \(v\), then \(W\) can be recovered as \(W = F(D \mid V)\), where \(F(d \mid v) := P(D \leq d \mid V = v)\). This can be consistently estimated from a sample, so \(W\) is "asymptotically" observable. A special case occurs when \(q\) is separable: \(D = q_0(V) + W\). This strictly exogenous instrument case thus covers a considerable range of the empirical work that does not assume strictly exogenous \(D\).

As Schennach, White, and Chalak (2011, prop. A.1) show, however, there are triangular structures where \(F(D \mid V)\) is not a valid covariate. This happens when \(q\) is not suitably monotone.
and/or $W$ is not scalar. With exogenous instruments $V$, certain marginal effects of $D$ on $Y$ can still be identified as a function of reduced form marginal effects of $V$ on $D$ and of $V$ on $Y$, using an extension of the method of indirect least squares, provided $q$ is separable. But when $q$ is nonseparable, the resulting effect biases cannot be eliminated.

Chalak and White (2011a) discuss numerous other structures where conditional exogeneity does not hold, but where other conditional independence assumptions can be used together with exclusion restrictions to identify effects of interest. Often, identification in such cases is achieved by exclusion restrictions that yield functions of effects identified by strict or conditional exogeneity.

Useful high level insight for specifying covariates arises from the fact that $D \perp U \mid X$ means that, given $X$, $U$ is of no value in predicting $D$, and vice versa. As the presence of $Z$ in $X$ is already dictated, this suggests that $W$ should be chosen so that, together with $Z$, $W$ predicts $U$ sufficiently well that there is nothing left for $D$ to predict, or vice versa. As White and Lu (2011a) show, the resulting estimator precision depends on whether $W$ acts as a predictor of $U$ or $D$. Choosing $W$ to predict $U$ results in more efficient estimators than choosing $W$ to predict $D$. Predicting $D$ too well results in the analog of extreme multicollinearity. Thus, one should seek covariates that proxy for $U$, the stronger the better, and avoid drivers of $D$ or proxies for these.

It may be thought that for $D \perp U \mid X$ to be plausible, one must have a large number of covariates, $X$. But as we saw above, $X$ is optional – identification can hold without any covariates. Plausibility is ensured by choosing $X$ to respect the necessary triangularity constraints and by providing further structural justifications for a specific choice of $X$.

Indeed, if proper care is not exercised, there are three reasons why having more covariates rather than fewer may be a bad idea. The first relates to Simpson’s paradox: even if $D \perp U \mid (Z,W_1)$ holds, there is no guarantee that $D \perp U \mid (Z,W_1,W_2)$ holds – adding covariates may destroy conditional exogeneity. Second, as just noted, even without violating B.2, adding covariates predicting $D$ can have adverse consequences for effect estimator efficiency. Finally, including certain covariates can result in identifying and estimating effects other than what one intended. Specifically, suppose that $D$ and $Z$ both drive $Y$, but that $D$ in fact causes $Z$, violating the triangularity restriction. If the researcher overlooks this and uses $Z$ as a covariate, then the direct effect of $D$ but not its full effect may be identified. To see this, consider the linear system

$$Y = D\beta_o + Z\gamma_o + U, \quad Z = D\alpha_o + V,$$

where $D \perp (U,V)$ and $(D,V) \perp U$. Then $(D,Z) \perp U$, so $(\beta_o,\gamma_o)$ is identified from the regression.
of \( Y \) on \( D \) and \( Z \). But the full effect of \( D \) is identified from the regression of \( Y \) on \( D \), omitting \( Z \):

\[
Y = D\beta_o + (D\alpha_o + V)\gamma_o + U = D(\beta_o + \alpha_o\gamma_o) + (U + V\gamma_o),
\]
as \( D \perp (U + V\gamma_o) \). The full effect of \( D \) on \( Y \) is \( \beta_o + \alpha_o\gamma_o \), which is the direct effect, \( \beta_o \), plus the indirect effect, \( \alpha_o\gamma_o \). Including \( Z \) leads to an "included variable" bias.

In specific contexts, the direct and/or indirect effect rather than the full effect may be of interest, so it is important to have means of recovering these. Imposing the Rosenbaum (1984) restriction that \( D \) does not drive the covariates, as in B.1, ensures identification of the full effect, but does not facilitate recovering direct or indirect effects. In contrast, the structural approach permits formulating a variant of B.1 where \( D \) affects \( Z \), facilitating recovery of direct and indirect effects. See Chalak and White (2011a,b) for further discussion of direct and indirect effects.

We complete this discussion by noting that regardless of the theoretical plausibility of conditional exogeneity for any given covariates, this need not be taken on faith. White and Chalak (2010) describe a variety of ways to test conditional exogeneity. See also White and Lu (2011b).

6 Identification and Identification Failure for General Effects

So far, the effects studied have been average effects, such as the average effect of treatment on the treated or the average marginal effect. Average effects are not the only relevant focus of interest, however. Attention has also been directed to the effects of treatment on aspects of the conditional response distribution other than the mean. Heckman, Smith, and Clements (1997) pursue this in the context of programme evaluation. For wage determination, Firpo, Fortin, and Lemieux (2009, 2011) study identification of effects of binary treatments on aspects of the unconditional response distribution, such as the variance, median, or density. Here, we extend this analysis by investigating identification and identification failure for the effects of both binary and continuous causes on the conditional response distribution. We consider effects on distributional aspects defined by optimality properties or by moment properties. These lead to plug-in effect estimators based on quasi-maximum likelihood or generalized method of moment estimators.

Although the average effect of treatment is also the effect of treatment on the average response, this equivalence does not hold for other aspects of the response distribution. Here, we examine the effect of a treatment on the response distribution, not the distribution of treatment effects. The latter requires a different analysis; see Fan and Park (2010) and the references given there.
6.1 Identifying Effects on Distributional Aspects

A comprehensive way to define distributional aspects of interest for a given response \( Y \) is to specify that these aspects satisfy some optimality property. Let \( Y \) be a scalar, and consider the \( \lambda \times 1 \) vector of distributional aspects \( \alpha \) that have the optimality property

\[
\alpha(d, x) := \arg \max_a \ E(\rho(Y, a) \mid D = d, X = x),
\]

where \( \rho : \mathbb{R} \times \mathbb{R}^\lambda \to \mathbb{R} \) is a researcher-specified optimality criterion.\(^{11}\) For example, let

\[
\rho_{mse}(y, a) = -0.5 \ (y - a)^2.
\]

Solving the first order conditions for the maximum with respect to \( a \) gives the conditional mean:

\[
\alpha(d, x) = E(Y \mid D = d, X = x).
\]

Alternatively, let \( \rho \) be the "check" function,

\[
\rho_{chk}(y, m) = -|y - a| \ (q \ 1[y \geq a] + (1 - q) \ 1[y < a]).
\]

With \( \rho = \rho_{chk} \), \( \alpha(d, x) \) is the conditional \( q \)-quantile.

In these examples, \( a \) is a scalar. With \( \lambda > 1 \), we can consider multiple distributional aspects. For example, let \( a := (a_1, a_2) \), and let \( \rho \) define the normal \( N(a_1, a_2) \) quasi-log-likelihood,

\[
\rho_{nml}(y, a) = -0.5 \ln a_2 - 0.5(y - a_1)^2/a_2.
\]

This focuses on the conditional mean and conditional variance of the distribution jointly, as

\[
\alpha_1(d, x) = E(Y \mid D = d, X = x) \quad \text{and} \quad \alpha_2(d, x) = \text{var}(Y \mid D = d, X = x),
\]

Instead of having an optimality property, suppose \( \alpha \) satisfies a generalized moment condition:

\[
E(\psi(Y, \alpha(d, x)) \mid D = d, X = x) = 0,
\]

where \( \psi : \mathbb{R} \times \mathbb{R}^\lambda \to \mathbb{R}^\kappa \), \( \kappa \geq \lambda \), is a researcher-specified moment function. For example, let \( \psi(y, a) = y - a \). Then we again obtain the conditional mean, \( \alpha(d, x) = E(Y \mid D = d, X = x) \). Alternatively, let \( \rho \) be an optimality criterion, as above, for which \( \psi = \nabla_a \rho \) exists. Provided \( \alpha(d, x) \)

\(^{11}\text{Our notation for } \rho \text{ and } \psi \text{ (below) follows that of Huber (1981). It would be more explicit to write } \alpha_{\varphi} \text{ here instead of } \alpha, \text{ but we leave the dependence on } \rho \text{ (or } \psi) \text{ understood to keep the notation simple. The same is true for } \varphi \text{ defined below.} \)
is an interior point, the necessary first-order conditions require $E(\nabla_a \rho(Y, \alpha(d, x)) \mid D = d, X = x) = 0$, so both the moment approach and the optimality approach define the same distributional aspects. The moment approach does not require $\psi = \nabla_a \rho$, so it is more comprehensive in this sense. On the other hand, the optimality approach can handle some cases, such as quantiles, that are not as easily treated by the moment approach. The two approaches are thus complementary.

To define general distributional $\alpha-$effects of treatment, let $\varphi_d(\delta, x)$ satisfy

$$\varphi_d(\delta, x) := \arg \max_a E(\rho(\delta, a) \mid D = d, X = x)$$

for the optimality approach or, for the moment approach,

$$E(\psi(\delta, \varphi_d(\delta, x)) \mid D = d, X = x) = 0.$$ 

We view $d$ as a given treatment value and $\delta$ as a counterfactual taking any admissible value. These can be vectors of discrete, categorical, and/or continuous variables. We give $\varphi_d(\delta, x)$ a structural interpretation below. When $\delta = d$, we have $\varphi_d(d, x) = \alpha(d, x)$. Analogs of all the treatment effects considered so far can be derived from the $\alpha-$effects of "discrete" treatments, $(d, \delta):$

$$\gamma_d(\delta, x) := \varphi_d(\delta, x) - \varphi_d(d, x).$$

With $d = 0$ and $\delta = 1$, we have $\gamma_0(1, x) = \varphi_0(1, x) - \varphi_0(0, x)$, the conditional $\alpha-$effect of binary treatment on the untreated. Taking $d = 1$ and $\delta = 0$ gives $\gamma_1(0, x) = \varphi_1(0, x) - \varphi_1(1, x)$, the conditional $\alpha-$effect of removing treatment from the treated, i.e., the opposite of the conditional $\alpha-$effect of treatment on the treated. The conditional marginal $\alpha-$effect of treatment is

$$\hat{\gamma}_d(x) := \nabla_\epsilon \gamma_d(d + \epsilon; x) \mid \epsilon = 0 = \nabla_\epsilon \varphi_d(d + \epsilon, x) \mid \epsilon = 0 = \nabla_\delta \varphi_d(d, x) \mid \delta = d.$$

To represent the effect bias, we write $\gamma_d(\delta, x) = [\varphi_d(\delta, x) - \varphi_d(d, x)] + [\varphi_d(\delta, x) - \varphi_d(\delta, x)]$, so

$$\alpha(\delta, x) = \alpha(d, x) + b_d(\delta, x),$$

where the discrete $\alpha-$effect bias is now $b_d(\delta, x) := \varphi_d(\delta, x) - \varphi_d(d, x)$, parallel to what we had earlier. Clearly, the necessary and sufficient condition for identification here is $b_d(\delta, x) = 0$. Either A.2 or B.2 suffices, as both imply $\varphi_d(\delta, x) = \varphi_d(d, x)$ in the first term of $b_d(\delta, x)$, for all $(d, \delta, x)$.

To give a formal result we use C.1 and C.2 of the Appendix. Assumption C.1 ensures that the needed moments exist. Assumption C.2 ensures that the needed $\varphi$’s exist.

**Theorem 6.1** Given B.1 and if C.1 and C.2 hold for $d, \delta$, and $x$ as specified below, then
(i) for \( x \in \mathcal{X} \) and \( d, \delta \in \text{int}(\mathcal{D}) \), \( b_d(\delta, x) = 0 \) if and only if
\[
\gamma_d(\delta, x) = \alpha(\delta, x) - \alpha(d, x);
\]
(ii) for \( d, \delta \in \text{int}(\mathcal{D}) \), (a) if \( b_d(\delta, x) = 0 \) holds for all \( x \) in \( \mathcal{X} \), then
\[
\tilde{\gamma}_d(\delta) = E(\alpha(\delta, X) - \alpha(d, X));
\]
(b) if \( \tilde{\gamma}_d(\delta) = E(\alpha(\delta, X) - \alpha(d, X)) \), then with \( \tilde{b}_d(\delta) := E(\alpha(\delta, X) - \alpha(d, X)) - \tilde{\gamma}_d(\delta) \),
\[
\tilde{b}_d(\delta) = E[1\{X \in \mathcal{X}_{d,\delta}\} b_d(\delta; X) \mid D = d] = 0,
\]
where \( \mathcal{X}_{d,\delta} := \{x \in \mathcal{X} : b_d(\delta, x) \neq 0\} \).

We omit stating the result for marginal effects; we take this up in the next section. A significant implication of this result is that either A.2 or B.2 suffices to identify all of these general effects globally, clearly demonstrating their power and convenience.

6.2 Identification Failure and Near Identification

To gain insight into identification failure beyond that provided by the necessary and sufficient identification conditions just given, we use the exogeneity score. For conciseness, we treat only the moment case. This also covers sufficiently regular optimization cases, those with \( \psi = \nabla_a \rho \).

First, we give \( \varphi_d(\delta, x) \) a structural representation. For this, define the non-negative weights\(^{12} \)
\[
\zeta_d(\delta, x, u) := dG(u \mid d, x) / dG(u \mid \delta, x).
\]
These integrate to one: \( \int \zeta_d(\delta, x, u) dG(u \mid \delta, x) = 1 \) for all \((d, \delta, x)\). Using these and B.1 gives
\[
E(\psi(Y_{\delta}, \varphi_d(\delta, x)) \mid D = d, X = x) = \int \psi(r(\delta, z, u), \varphi_d(\delta, x)) dG(u \mid d, x)
\]
\[
= \int \psi(r(\delta, z, u), \varphi_d(\delta, x)) \zeta_d(\delta, x, u) dG(u \mid \delta, x)
\]
\[
= \int \psi(r, \varphi_d) \zeta_d dG = 0.
\]
In the last expression we begin temporarily suppressing \( u, \delta, \) and \( x \) to simplify notation. We see that \( \varphi_d \) is a weighted moment, with weights reflecting deviations from conditional exogeneity.

To examine \( b_d = \varphi_d - \varphi_d^\prime \), we use a Taylor expansion about \( \varphi_d^\prime \). With sufficient regularity,
\[
\int \psi(r, \varphi_d) \zeta_d dG = \int \psi(r, \varphi_d) \zeta_d dG + \{\nabla_{\alpha'} \int \psi(r, \varphi_d) \zeta_d dG\} [\varphi_d - \varphi_d^\prime] + o(1).
\]

\(^{12}\)We implicitly assume \( U_{\delta, x} = U_{d, x} \), which avoids division by zero and ensures integration to one.
As $\int \psi(r, \varphi_d) \zeta_d \, dG = 0$ and $\int \psi(r, \varphi_\delta) \, dG = 0$, we have
\[
\{\nabla_a \int \psi(r, \varphi_d) \zeta_d \, dG\} [\varphi_d - \varphi_\delta] + o(1) = -\int \psi(r, \varphi_\delta) \zeta_d \, dG = \int \psi(r, \varphi_\delta) \left(1 - \zeta_d\right) \, dG.
\]
Letting $s_d := 1 - \zeta_d$ and defining the $\kappa \times \lambda$ matrix $J_d := -\nabla_a \int \psi(r, \varphi_\delta) \zeta_d \, dG$, we have
\[
J_d b_d = \int \psi(r, \varphi_\delta) \, s_d \, dG + o(1).
\]
When $J_d$ has full rank $\lambda$, let $K_d := [J_d' J_d]^{-1} J_d'$. Solving for $b_d$ gives
\[
b_d = K_d \int \psi(r, \varphi_\delta) \, s_d \, dG + o(1) \quad \text{or} \quad b_d(\delta, x) = K_d(\delta, x) E(\varepsilon S_{d,D} \mid D = \delta, X = x) + o(1).
\]
This is the generalized omitted variable bias representation, viewing the exogeneity score $S_{d,D} := 1 - \zeta_d(D, X, U)$ as the omitted variable and taking the "generalized regression errors" to be the $\kappa \times 1$ vector $\varepsilon := \psi(Y, \alpha(D, X))$. (If $\psi(y, a) = y - a$, then $\varepsilon = Y - E(Y \mid D, X)$.)

From this, we see that, apart from $o(1)$ terms, the necessary and sufficient condition for the discrete moment effect $\gamma_d(\delta, x)$ to be identified as $\alpha(\delta, x) - \alpha(d, x)$ is that
\[
J_d(\delta, x)' E(\varepsilon S_{d,D} \mid D = \delta, X = x) = 0. \tag{2}
\]
Near identification results obtain by applying, for example, a matrix version of Cauchy-Schwarz. For this, we define $\sigma^2_d(\delta, x) := E(\varepsilon' \varepsilon \mid D = \delta, X = x)$ and $\sigma^2_{S_{d,D}}(\delta, x) := E(S^2_{d,D} \mid D = \delta, X = x)$.

When $\kappa = \lambda$, the necessary and sufficient identification condition is simply
\[
E(\varepsilon S_{d,D} \mid D = \delta, X = x) = 0.
\]
When $\kappa > \lambda$, a further "lucky" opportunity for identification occurs when the covariance in (2) is in the null space of $J_d(\delta, x)'$. But this is not a viable foundation for local identification.

For the Taylor expansion to work, $\varphi_d(\delta, x)$ must be close enough to $\varphi_\delta(\delta, x)$ to make the higher order terms small. For this, $G(\cdot \mid \delta, x)$ has to be close to $G(\cdot \mid d, x)$. Theorem A.1 of the Appendix gives conditions ensuring the substantive result that $\varphi_d(\delta, x)$ depends continuously on the exogeneity score. Imposing C.3 of the Appendix ensures this, so when $\sigma^2_{S_{d,D}}(\delta, x)$ is small, then $\varphi_d(\delta, x)$ is close to $\varphi_\delta(\delta, x)$. Note that $\delta$ need not be close to $d$ for this. C.4 of the Appendix imposes the other conditions for the Taylor expansion. For discrete $\alpha$-effects, we have

**Theorem 6.2** Given Assumptions B.1 and C.1-C.4, suppose that for given $x \in X$ and $d, \delta \in \text{int}(D)$, $\sigma_{S_{d,D}}(\delta, x) < \eta_c$, where $\epsilon > 0$ is given in C.4 and $\eta_c$ is sufficiently small. Then
\[
b_d(\delta, x) = K_d(\delta, x) E(\varepsilon S_{d,D} \mid D = \delta, X = x) + o(1) \quad \text{and}
\]
The result for marginal \( \alpha \)-effects is similar. Here, we have \( \tilde{\alpha}_d(x) = \tilde{\gamma}_d(x) + \tilde{b}_d(x) \), where \( \tilde{\alpha}_d(x) := \nabla_{\delta} \alpha(d, x) \) and \( \tilde{b}_d(x) := \nabla_{\delta} b(\delta, x) \mid_{\delta=d} \). The necessary and sufficient condition for identification of the marginal \( \alpha \)-effect, \( \tilde{\gamma}_d(x) \), is therefore \( \tilde{b}_d(x) = 0 \). Our next result gives the exogeneity score/omitted variable bias representation for \( \tilde{b}_d(x) \) and a near identification result.

Regularity conditions and derivations are in the Appendix. To state the result, we define \( \tilde{J}_d = - \int \nabla_{\alpha} \psi(r, \alpha) \, dG = - \int \nabla_{\alpha} \psi(r(d, z, u), \alpha(d, x)) \, dG(u \mid d, x) =: \tilde{J}_d(x) \) and \( \tilde{K}_d := [\tilde{J}_d \tilde{J}_d^{-1}]^{-1} \tilde{J}_d' \). Again, \( \varepsilon := \psi(Y, \alpha(D, X)) \), and the relevant exogeneity score is \( S = \nabla_d \ln g(U \mid D, X) \), with \( d \) a scalar for simplicity.

**Theorem 6.3** Given B.1 and B.4, suppose C.5 and C.6 hold for given \( x \in \mathcal{X} \) and \( d \in \text{int}(D) \), with \( \tilde{J}_d(x) \) having full rank \( \lambda \). Then

\[
\tilde{b}_d(x) = \tilde{K}_d(x) \, E(\varepsilon S \mid D = d, X = x).
\]

If \( \sigma_s(\delta, x) < \infty \) and \( \sigma_S(\delta, x) < \infty \), then

\[
[b_d(\delta, x)' J_d(\delta, x)' J_d(\delta, x) b_d(\delta, x)]^{1/2} \leq \lambda^{1/2} \sigma_{s}(\delta, x) \sigma_{S}(d, x).
\]

We have \( \nabla_d \ln g(u \mid d, x) = \nabla_{\varepsilon} [1 - g(u, d + \varepsilon, x)] \mid_{\varepsilon=0} \), so the marginal exogeneity score is the derivative of the discrete exogeneity score. When \( \psi(y, \alpha) = y - a \), we obtain the results of Section 4. We omit results for the other effects, as these are parallel to the results of Section 4.

## 7 Estimating Effects

To estimate any of the treatment effects discussed in previous sections, it suffices to suitably estimate \( \alpha(d, x) \), as these effects are either differences or derivatives involving \( \alpha \). The plug-in principle again applies, once we specify a suitable model for \( \alpha \). With this in place, estimation is completely standard, so we only sketch the details and give pertinent references.

A model \( \mathcal{M} \) is a collection of functions \( m : D \times \mathcal{X} \times \Theta \to \mathbb{R} \), where \( \Theta \) is a parameter space:

\[
\mathcal{M} := \{ m(\cdot, \cdot ; \theta) : D \times \mathcal{X} \to \mathbb{R} \mid \theta \in \Theta \}.
\]

A linear model occurs when \( m(d, x, \theta) = d' \beta + x' \pi \), with \( \theta := (\beta, \pi) \). Alternatively, \( \theta \) could be a function: \( m(d, x, \theta) = \theta(d, x) \). In either case, \( \theta \) is a parameter, but we follow common usage by referring to cases with finite dimensional \( \Theta \) as "parametric" and to cases with infinite dimensional
\( \Theta \) as "nonparametric." In some applications, \( \Theta \) may have the form \( \Theta = \Theta_1 \times \Theta_2 \), with \( \Theta_1 \) finite dimensional and \( \Theta_2 \) infinite dimensional. Such cases are commonly called "semi-parametric."

For simplicity, we assume the model is correctly specified: there exists \( \theta_o \) in \( \Theta \) such that \( m(\cdot, \cdot; \theta_o) = \alpha \). For the optimality approach, this implies that

\[
\theta_o := \arg \max_{\theta \in \Theta} E[\rho(Y, m(D, X, \theta))].
\]

For the moment approach, \( \theta_o \) satisfies

\[
E[\psi(Y, m(D, X, \theta_o))] = 0.
\]

For simplicity, we assume that \( \theta_o \) is unique.\(^{13}\)

Estimating \( \theta_o \) is now completely standard. Applying the plug-in principle, we replace expectations with averages from a sample \( (Y_i, D_i, X_i)_{i=1}^n \), yielding estimators \( \hat{\theta}_n \) of \( \theta_o \) as

\[
\hat{\theta}_n := \arg \max_{\theta \in \Theta_n} n^{-1} \sum_{i=1}^n \rho(Y_i, m(D_i, X_i, \theta)) \quad \text{or} \quad n^{-1} \sum_{i=1}^n \psi(Y_i, m(D_i, X_i, \hat{\theta}_n)) = o_P(1).
\]

When \( \Theta \) is finite dimensional, we can take \( \Theta_n = \Theta \) for all \( n \). Huber (1981) and White (1994) give conditions ensuring the consistency and asymptotic normality of these estimators. When \( \Theta \) is infinite dimensional, we can apply the method of sieves (e.g., Grenander, 1981; Chen, 2007), with \( \text{cl}(\cup \Theta_n) = \Theta \). For the moment case, the (sieve) GMM estimator satisfies

\[
\arg \min_{\theta \in \Theta_n} [n^{-1} \sum_{i=1}^n \psi(Y_i, m(D_i, X_i, \theta))]' \hat{V}_n^{-1} [n^{-1} \sum_{i=1}^n \psi(Y_i, m(D_i, X_i, \theta))],
\]

where \( \hat{V}_n \) is consistent for positive definite \( V \), e.g., \( V := E[\psi(Y, m(D, X, \theta_o)) \psi(Y, m(D, X, \theta_o))]' \). See Hansen (1982) and Ai and Chen (2003) for general treatments. One can also apply Owen’s (1988, 2001) empirical likelihood methods (see also, e.g., Schennach (2007) or Ragusa (2011)).

When nonparametric or semiparametric methods are used to estimate marginal effects, \( m \) and \( \{\Theta_n\} \) must be properly chosen to ensure that \( \nabla_d \hat{m}_n := \nabla_d m(\cdot, \cdot; \hat{\theta}_n) \) is consistent for \( \nabla_d \alpha \).

Kernel methods apply straightforwardly to nonparametric estimation of explicit moments, \( E[\tau(Y) \mid D, X] \) (here, \( \psi(y, a) = \tau(y) - a \)). See, e.g., Pagan and Ullah (1999), Li and Racine (2007). Li, Lu, and Ullah (2003), and Schennach, White, and Chalak (2011) give results for estimating derivatives of conditional expectations.

\(^{13}\)This uniqueness is a different type of identification condition, distinct from that of Hurwicz (1950). See, e.g., White (1994, pp.28-29). To distinguish the two concepts, we may call this uniqueness "stochastic identification" and that of Hurwicz "structural identification". Neither is necessary for the other.
To form estimators of treatment-specific $\alpha$-effects, simply average over the appropriate subpopulation. For example, the binary $\alpha$-effect of treatment on the treated can be estimated as

$$\hat{\gamma}_{1,n} = n_1^{-1} \sum_{i:D_i = 1} [m(1, X_i; \hat{\theta}_n) - m(0, X_i; \hat{\theta}_n)],$$

where $n_1$ is the number of observations with $D_i = 1$.

In Section 3, we identified certain effects from a system of simultaneous equations as

$$\nabla_{y_2} s_1(r_2(d, z, u), d, z, u) = [\nabla_{d_2} r_2(d, z, u)' \nabla_{d_2} r_2(d, z, u)]^{-1} \nabla_{d_2} r_2(d, z, u)' \nabla_{d_2} r_1(d, z, u).$$

In Section 4.1, we saw that in the separable case, $r(D, Z, U) = r_0(D, Z) + U$, we have $\nabla_d r_0(d, z) = \nabla_d r(d, z, u)$. When $r_1$ and $r_2$ are separable, with the obvious notation we have

$$\nabla_{y_2} s_1(r_2(d, z, u), d, z, u) = [\nabla_{d_2} r_{2,0}(d, z)' \nabla_{d_2} r_{2,0}(d, z)]^{-1} \nabla_{d_2} r_{2,0}(d, z)' \nabla_{d_2} r_{1,0}(d, z).$$

Suppose we estimate $\nabla_{d_2} r_{1,0}$ and $\nabla_{d_2} r_{2,0}$ consistently using any method described above, as, say, $\nabla_{d_2} \hat{r}_{1,n}$ and $\nabla_{d_2} \hat{r}_{2,n}$. Then a consistent estimator of $\nabla_{y_2} s_1(Y_{2,i}, D_i, Z_i, U_i)$ is

$$\nabla_{y_2} \hat{s}_{1,n}(Y_{2,i}, D_i, Z_i, U_i) = [\nabla_{d_2} \hat{r}_{2,n}(D_i, Z_i)' \nabla_{d_2} \hat{r}_{2,n}(D_i, Z_i)]^{-1} \nabla_{d_2} \hat{r}_{2,n}(D_i, Z_i)' \nabla_{d_2} \hat{r}_{1,n}(D_i, Z_i).$$

This is an analog of Haavelmo’s (1943) ILS method for linear structural systems. This analog applies to nonlinear systems, and it may be parametric, semiparametric, or nonparametric. In contrast to standard ILS, which uses classical exogenous instruments ($D \perp U$), this analog uses conditionally exogenous instruments ($D \perp U \mid X$) to construct $\nabla_{d_2} \hat{r}_{1,n}$ and $\nabla_{d_2} \hat{r}_{2,n}$.

Asymptotic distribution results for such estimators can typically be derived straightforwardly, for example by applying the delta method to the asymptotic joint distribution of $(\nabla_{d_2} \hat{r}_{1,n}, \nabla_{d_2} \hat{r}_{2,n})$. As the details are involved, we leave this to future work.

Although separability for $r_2$ is crucial for this method, separability for $r_1$ can be dropped, and one can still identify and consistently estimate the covariate-conditioned average effect

$$E(\nabla_{y_2} s_1(Y_2, D, Z, U) \mid D, X).$$

This application of treatment effect estimation methods to systems satisfying simultaneous equations appears to represent a new approach to identifying and estimating such structural effects (see Angrist, Graddy, and Imbens (2000) for related results using exogenous instruments).
8 Conditional Exogeneity or Weak Unconfoundedness?

As we have seen, neither weak unconfoundedness (A.2) nor conditional exogeneity (B.2) is necessary to identify the effects considered here. On the other hand, although the various necessary and sufficient identification conditions have a similar form, they also vary from case to case. Both weak unconfoundedness and conditional exogeneity have the considerable virtue of ensuring global identification for all the effects considered here. If we limit ourselves to just these two choices, which should we prefer? We now discuss some properties that may help in deciding.

Recall that Proposition 4.1 established that B.2 implies A.2. Does A.2 suffice for B.2? The answer is negative but nuanced, as the following results show.

**Proposition 8.1** Suppose that B.1 holds and that there exist $d \in \mathcal{D}$ and a function $f$ such that $U = f(Y_d, X)$. If $Y_d \perp D \mid X$, then B.2 holds.

It suffices that $U = f(Y_d, X)$ and $Y_d \perp D \mid X$ for a single $d$. Thus, we only use part of A.2, but we also impose certain structure on $r$. When A.2 holds and $r$ is separable for some (not all) $d$, then B.2 holds, as separability ensures $U = f(Y_d, X) = Y_d - r_0(d, Z)$. More generally, $f$ exists when $r$ is invertible for some $d$, as $U = r^{-1}(d, Z, Y_d)$. Monotonicity of $r(d, z, \cdot)$ for some $d$ and all $z$ insures invertibility but is not necessary.

If only part of A.2 is used in Proposition 8.1, can we use more of A.2 and eliminate the need for $f$? The answer is negative. Nevertheless, the next result shows that we can weaken the requirements on $f$ by using more of A.2, together with some further unconfoundedness requirements.

**Proposition 8.2** Suppose that B.1 holds, and that there exist a non-empty subset $\mathcal{D}^*$ of $\mathcal{D}$ and a function $f$ such that $U = f(\{Y_d, d \in \mathcal{D}^*\}, X)$. If $\{Y_d, d \in \mathcal{D}^*\} \perp D \mid X$, then B.2 holds.

This gives Proposition 8.1 as a corollary. The condition $\{Y_d, d \in \mathcal{D}^*\} \perp D \mid X$ is a "moderate" unconfoundedness assumption weaker than standard unconfoundedness ($\{Y_d, d \in \mathcal{D}\} \perp D \mid X$), but neither necessary nor sufficient for A.2. Moderate unconfoundedness is equivalent to either (i) $Y_d \perp D \mid X$ and $\{Y_{d^*}, d^* \in \mathcal{D}^* \setminus \{d\} \} \perp D \mid (X, Y_d)$ for each $d \in \mathcal{D}^*$; or (ii) $\{Y_{d^*}, d^* \in \mathcal{D}^* \setminus \{d\} \} \perp D \mid X$ and $Y_d \perp D \mid (X, \{Y_{d^*}, d^* \in \mathcal{D}^* \setminus \{d\} \})$ for each $d \in \mathcal{D}^*$, by lemma 4.3 of Dawid (1979).

Further results are possible that take greater advantage of the presence of $X$, given additional structure. We leave these aside here, however, as Propositions 4.1, 8.1, and 8.2 make the situation clear: A.2 is weaker than B.2, but it is also complementary, in the sense that it is not necessary for
B.2 in the presence of other commonly assumed structure. Also, the fact that random treatment or other underlying structure (cited in Section 5) can justify B.2 and the fact that B.2 is explicit about economic structure and can be testable may give conditional exogeneity certain appeal over weak unconfoundedness. But either works for identification of causal effects, and the choice between them can be a matter of taste.

9 Summary and Conclusion

In this paper, we study a system of structural equations encompassing the systems commonly used in econometrics, including triangular systems compatible with simultaneous equations, and we provide conditions under which this system is also formally equivalent to the data generating process underlying the potential outcome approach to estimating treatment effects. This permits economic theory to be built into treatment effect methods. We characterize the limits to effect identification in either approach by providing necessary and sufficient conditions for identification, elucidating the sources and consequences of identification failure by examining the biases arising when the necessary conditions fail. The exogeneity score emerges as a central quantity, governing the magnitude of the effect bias and the possibilities for near identification, and affording an omitted variable representation for effect biases. We also examine the relations between unconfoundedness, conditional exogeneity, and the necessary and sufficient identification conditions. The necessary and sufficient conditions vary from case to case, but weak unconfoundedness and conditional exogeneity ensure global identification not only of average effects of treatment but also of effects on any aspect of the response distribution represented as an optimality property or generalized moment of the conditional response distribution.

Our analysis elucidates insights that each of the treatment effect/potential outcome and the structural approaches offers when viewed in terms of the other. In particular, blending these approaches appears to offer the opportunity for useful advances in analyzing efficiency bounds for extremum and moment-based estimators of effects, in developing further tests for unconfoundedness or conditional exogeneity, and in developing tests for restrictions, such as monotonicity or separability, on otherwise general nonseparable structures.

A Mathematical Appendix

For convenience, we refer to Dawid (1979) as D79 in what follows.
Proof of Proposition 2.1: The proof is direct and is omitted. ■

Proof of Proposition 4.1: By lemmas 4.1 and 4.2(i) of D79, B.1 and $V \perp U \mid X$ imply $(X, V) \perp U \mid X$, which then implies $D \perp U \mid X$. It follows from lemmas 4.1 and 4.2(i) of D79 that $U \perp D \mid X$ implies $(U, X) \perp D \mid X$. Lemma 4.2(ii) of D79 then gives $\{Y_d, d \in D\} \perp D \mid X$, with $Y_d = r(d, Z, U)$. This implies A.2 by D79, lemma 4.3. ■

Proof of Theorem 4.2: The proof of (i) is given in the text, with B.3 justifying the interchange of integral and derivative, using the mean value theorem and the Lebesgue dominated convergence theorem (e.g., Bartle, 1966, corollary 5.9). The easy proofs of (ii) and (iii) are omitted. ■

For Assumptions B.4 - B.7, we let $d \in \text{int}(D)$ and $x := (w, z) \in X$ be given, and we let $C$ be a non-empty open subset of $D \subseteq \mathbb{R}$ containing $d$.

Assumption B.4 There is a $\sigma$-finite measure $\nu(\cdot \mid d, x)$ such that for all $\epsilon$ such that $d + \epsilon \in C$, $G(\cdot \mid d + \epsilon, x) \ll \nu(\cdot \mid d, x)$.

Assumption B.5 For all $\epsilon$ such that $d + \epsilon \in C$: (i.a) $E(r(d + \epsilon, Z, U) \mid D = d + \epsilon, X = x) < \infty$ and (i.b) $E(r(d + \epsilon, Z, U) \mid D = d, X = x) < \infty$; and (ii) $\nabla_\epsilon r(d + \epsilon, z, u)$ exists for almost every $u$ in $U_{d,x}$. (iii) There is a random variable $\Delta_{d,z}(U)$ with $E(\Delta_{d,z}(U) \mid D = d, X = x) < \infty$ such that for all $\epsilon$ such that $d + \epsilon \in C$, $|\nabla_\epsilon r(d + \epsilon, z, u)| \leq \Delta_{d,z}(u)$ for almost every $u$ in $U_{d,x}$.

Assumption B.6 For all $\epsilon$ such that $d + \epsilon \in C$: (i) $E(r(d + \epsilon, Z, U) \mid [g(U \mid d + \epsilon, x)/g(U \mid d, x) - 1] \mid D = d, X = x) < \infty$; and (ii) $\nabla_\epsilon g(u \mid d + \epsilon, x)$ exists for almost every $u$ in $U_{d,x}$. (ii) There is a random variable $\Delta_{d,x}(U)$ with $E(\Delta_{d,x}(U) \mid D = d, X = x) < \infty$ such that for all $\epsilon$ such that $d + \epsilon \in C$, $|r(d + \epsilon, z, u) \nabla_\epsilon \ln g(u \mid d + \epsilon, x)| \leq \Delta_{d,x}(u)$ for almost every $u$ in $U_{d,x}$.

Proof of Theorem 4.3: (i) The result follows by showing

$$
\nabla_d E(Y \mid D = d, X = x) - \tilde{\gamma}_d(x) = \int_{U_{d,x}} r(d, z, u) \nabla_d \ln g(u \mid d, x) dG(u \mid d, x).
$$

For all $\epsilon$ sufficiently small, $d + \epsilon \in C$, so

$$
\nabla_d E(Y \mid D = d, X = x) - \tilde{\gamma}_d(x) = \nabla_d \int_{U_{d,x}} r(d, z, u) g(u \mid d, x) d\nu(u \mid d, x) - \tilde{\gamma}_d(x)
$$

by B.1, B.4, and B.5(i.a)

$$
= \lim_{\epsilon \to 0} \epsilon^{-1} \int_{U_{d,x}} r(d + \epsilon, z, u) g(u \mid d + \epsilon, x) d\nu(u \mid d, x) - \int_{U_{d,x}} r(d, z, u) g(u \mid d, x) d\nu(u \mid d, x)
$$

by B.4 and B.5(i.a)

$$
- \tilde{\gamma}_d(x)
$$

33
\[= \lim_{\epsilon \to 0} \epsilon^{-1} \int_{U_{d,x}} r(d + \epsilon, z, u) g(u \mid d + \epsilon, x) \, dv(u \mid d, x) - \int_{U_{d,x}} r(d + \epsilon, z, u) g(u \mid d, x) \, dv(u \mid d, x)\]

\[+ \lim_{\epsilon \to 0} \epsilon^{-1} \int_{U_{d,x}} (r(d + \epsilon, z, u) - r(d, z, u)) g(u \mid d, x) \, dv(u \mid d, x) - \gamma_d(x) \quad \text{by B.5(i, b)}\]

\[= \lim_{\epsilon \to 0} \epsilon^{-1} \int_{U_{d,x}} r(d + \epsilon, z, u) g(u \mid d + \epsilon, x) \, dv(u \mid d, x) - \int_{U_{d,x}} r(d + \epsilon, z, u) g(u \mid d, x) \, dv(u \mid d, x) \quad \text{by B.5}.\]

Collecting terms, we have

\[\nabla_d E(Y \mid D = d, X = x) - \gamma_d(x)\]

\[= \lim_{\epsilon \to 0} \epsilon^{-1} \int_{U_{d,x}} r(d + \epsilon, z, u) \left[\frac{g(u \mid d + \epsilon, x) - g(u \mid d, x)}{g(u \mid d, x)}\right] g(u \mid d, x) \, dv(u \mid d, x)\]

\[= \int_{U_{d,x}} r(d, z, u) \lim_{\epsilon \to 0} \epsilon^{-1} \left[\frac{g(u \mid d + \epsilon, x) - g(u \mid d, x)}{g(u \mid d, x)}\right] g(u \mid d, x) \, dv(u \mid d, x) \quad \text{by B.6}\]

\[= \int_{U_{d,x}} r(d, z, u) \nabla_d \ln g(u \mid d, x) g(u \mid d, x) \, dv(u \mid d, x),\]

as was to be shown.

(ii) The result for (a) is obvious. To show (b), we write

\[\tilde{b}_d = E[\nabla_d E(Y \mid D = d, X)] - \gamma_d = E[\tilde{\gamma}_d(X) + \tilde{b}_d(X) \mid D = d] - \gamma_d\]

\[= E[\tilde{b}_d(X) \mid D = d]\]

\[= E[\{X \notin \mathcal{X}_d\} \tilde{b}_d(X) \mid D = d] + E[\{X \in \mathcal{X}_d\} \tilde{b}_d(X) \mid D = d]\]

\[= E[\{X \in \mathcal{X}_d\} \tilde{b}_d(X) \mid D = d] = 0,\]

where the final equality holds because \(E[\nabla_d E(Y \mid D = d, X)] - \gamma_d = 0.\]

**Assumption B.7** There is a measurable \(M_{d,x} : U_{d,x} \to \mathbb{R}^+\) with \(\int_{U_{d,x}} M_{d,x}(u) \, dv(u \mid d, x) < \infty\) such that for all \(\epsilon\) such that \(d + \epsilon \in C\), \(|\nabla \epsilon g(u \mid d + \epsilon, x)| \leq M_{d,x}(u)\) for almost every \(u \in U_{d,x}\).

Since \(\int_{U_{d,x}} g(u \mid d, x) \, dv(u \mid d, x) = 1\), it follows from B.6(ii) and B.7 that \(\int_{U_{d,x}} g(u \mid d + \epsilon, x) \, dv(u \mid d, x) < \infty\) for all \(\epsilon\) such that \(d + \epsilon \in C\). (Apply the mean value theorem given B.6(ii), the triangle inequality, and the domination condition of B.7.)

**Proof of Corollary 4.4:** This is given in the text, except for showing that \(\int_{U_{d,x}} [\nabla_d g(u \mid d, x)] \, dv(u \mid d, x) = \nabla_d \int_{U_{d,x}} g(u \mid d, x) \, dv(u \mid d, x)\). We have

\[\int_{U_{d,x}} [\nabla_d g(u \mid d, x)] \, dv(u \mid d, x) = \int_{U_{d,x}} \nabla \epsilon g(u \mid d + \epsilon, x) \, dv(u \mid d, x)\]
Proof of Proposition 4.5: (i) The result follows if \( E(Y \mid D = 1, X = x) - E(Y \mid D = 0, X = x) - \gamma_1^*(x) = E[Y_0 \mid D = 1, X = x] - E[Y_0 \mid D = 0, X = x] \), as \( E[Y_0 \mid D = 1, X = x] - E[Y_0 \mid D = 0, X = x] = 0 \) is equivalent to \( E[Y_0 \mid D, X = x] = E[Y_0 \mid X = x] \). Now

\[
E(Y \mid D = 1, X = x) - E(Y \mid D = 0, X = x) - \gamma_1^*(x) = E(Y_1 \mid D = 1, X = x) - E(Y_0 \mid D = 0, X = x) - \{E(Y_1 \mid D = 1, X = x) - E(Y_0 \mid D = 1, X = x)\}
= E(Y_0 \mid D = 1, X = x) - E(Y_0 \mid D = 0, X = x).
\]

(ii) The result for (a) is obvious. To show (b), we write

\[
b_1 = E(Y \mid D = 1) - \mu_{0,1} - \gamma_1 = E(b_1^*(X) \mid D = 1)
= E[1\{X \notin \mathcal{X}_1\} b_1^*(X) \mid D = 1] + E[1\{X \in \mathcal{X}_1\} b_1^*(X) \mid D = 1]
= E[1\{X \in \mathcal{X}_1\} b_1^*(X) \mid D = 1] = 0,
\]

where the last equality follows since \( E(Y \mid D = 1) - \mu_{0,1} = \gamma_1 \). ■

For the next results, let \( \mathcal{Y} := \text{supp}(Y) \) and let \( A \subset \mathbb{R}^\lambda, \lambda \in \mathbb{N} \), have non-empty interior. We let \( \mathcal{B}_Y \) denote the Borel \( \sigma \)-field generated by \( Y \), \( \mathcal{B}_A \) denote the Borel \( \sigma \)-field generated by the open sets of \( A \), and \( \mathcal{B}^\kappa, \kappa \in \mathbb{N} \), denote the Borel \( \sigma \)-field generated by the open sets of \( \mathbb{R}^\kappa \). We write the product \( \sigma \)-field for \( \mathcal{Y} \times A \) as \( \mathcal{B}_Y \times \mathcal{B}_A \). We say that \( \psi : \mathcal{Y} \times A \to \mathbb{R}^\kappa \) is jointly measurable if it is measurable\(\mathcal{B}_Y \times \mathcal{B}_A / \mathcal{B}^\kappa \), and similarly for \( \rho \).

Assumption C.1 Either (a) \( \rho : \mathcal{Y} \times A \to \mathbb{R} \) is jointly measurable and \( E[|\rho(Y, a)|] < \infty \) for all \( a \in A \); or (a) \( \psi : \mathcal{Y} \times A \to \mathbb{R}^\kappa \) is jointly measurable and \( E[|\psi(Y, a)^\prime \psi(Y, a)|^{1/2}] < \infty \) for all \( a \in A \).

Assumption C.2 \( \rho \) or \( \psi \) and, for given \( x \in \mathcal{X} \) and \( d, \delta \in \mathcal{D}, G(\cdot \mid D = d, X = x) \) and \( G(\cdot \mid D = \delta, X = x) \), are such that the equations defining \( \varphi_{\delta}(d, x) \), \( \varphi_{\delta}(\delta, x) \), and \( \varphi_d(\delta, x) \) each
have at least one solution in $A$. If there are multiple solutions, $\varphi_d(d,x)$, $\varphi_\delta(\delta,x)$, and $\varphi_\delta(\delta,x)$ represent a measurable selection from the solution set.

Assumption C.2 is a high-level assumption. Primitive conditions ensuring this follow from the theorem of the maximum (e.g., theorem 6.1.31 of Corbae, Stinchcombe, and Zeman, 2009, "CSZ") for $\rho$ or, e.g., the implicit function theorem (e.g., CSZ, pp. 322-326) for $\psi$.

**Proof of Theorem 6.1**: Assumptions B.1, C.1, and C.2 suffice for the derivations of the text.

To prove Theorem 6.2, we use the following result of Stinchcombe (2011):

**Theorem A.1**: Let $(\Omega, \mathcal{F})$ be a non-empty set and a $\sigma$-field of subsets, $\nu$ a $\sigma$-finite, countably additive, non-negative measure on $\mathcal{F}$, $A$ a compact metric space, $p > 1$, and $1/p + 1/q = 1$. Suppose $\psi : \Omega \times A \to \mathbb{R}$ is jointly measurable and $\psi(\omega, \cdot)$ is continuous a.e. $- \nu$; and $a \mapsto ||\psi(\cdot, a)||_q := \{ \int |\psi(\cdot, a)|^q d\nu \}^{1/q}$ is a continuous function from $A$ to $L^q := \{ h : ||h||_q < \infty \}$. Define the correspondence $\Gamma : L^p \Rightarrow A$ as $\Gamma(\zeta) := \{ a \in A : \int \psi(\cdot, a) \zeta d\nu = 0 \}$. Then $\Gamma$ is upper hemi-continuous (uhc).

**Proof**: Since $a \mapsto ||\psi(\cdot, a)||_q$ is continuous, the mapping $(a, \zeta) \mapsto f(a, \zeta) := \int \psi(\cdot, a) \zeta d\nu$ is jointly continuous (by Hölder’s inequality). Thus, the graph of $\Gamma$ is the closed set $f^{-1}(0)$. From corollary 6.1.33 of Corbae, et al. (2009, CSZ), $\Gamma$ is uhc; and as $\Gamma^*(L^*)$ has compact closure, $\Gamma^*$ is uhc. By theorem 6.1.23 of CSZ, a uhc singleton-valued correspondence is a continuous function.

**Assumption C.3** (i) $A$ is compact. For $x \in X$ and $\delta \in D : \psi(r(\delta, z, u), \cdot)$ is continuous a.e. $- G(\cdot | \delta, x); a \mapsto ||\psi(r(\delta, z, \cdot), a)||_2 := \{ \int |\psi(r(\delta, z, \cdot), a)|^2 dG(u | \delta, x) \}^{1/2}$ is continuous; and let

$$Z_{\delta, x} := \{ \zeta_d(\delta, x, \cdot), d \in D : \zeta_d(\delta, x, \cdot) \geq 0, \int_{U_{\delta, x}} \zeta_d(\delta, x, u) dG(u | \delta, x) = 1, \int_{U_{\delta, x}} |\zeta_d(\delta, x, u)|^2 dG(u | \delta, x) \}^{1/2} < \infty, \quad \zeta_\delta(\delta, x, \cdot) = 1 \}.$$

For $\zeta(\delta, x, \cdot)$ in $Z_{\delta, x}$, define

$$\Gamma_{\delta, x}(\zeta(\delta, x, \cdot)) := \{ a \in A : \int_{U_{\delta, x}} \psi(r(\delta, z, u), a) \zeta(\delta, x, u) dG(u | \delta, x) = 0 \}.$$

---

14 Upper hemi-continuity is the analog for correspondences of upper semi-continuity for functions. See CSZ, pp.150-151.
Let $Z^*_{\delta,x} \subset Z_{\delta,x}$ be such that for each $\zeta$ in $Z^*_{\delta,x}$, $\Gamma_{\delta,x}(\zeta(\delta, x, \cdot))$ has exactly one element.

(ii) For given $x \in \mathcal{X}$ and $\delta, d \in \mathcal{D}$, $\varphi_d(\delta, x)$ and $\varphi_d(\delta, x)$ are unique.

**Assumption C.4** For $x \in \mathcal{X}$ and $d, \delta \in \mathcal{D}$: (i) $E(\psi(Y_\delta, a) \mid D = d, X = x) < \infty$ for all $a \in A$; (ii) each element of $E(\psi(Y_\delta, \cdot) \mid D = d, X = x) : A \rightarrow \mathbb{R}^\kappa$ is differentiable in an open $\epsilon$–ball around $a = \varphi_d(\delta, x)$; (iii) $J_d(\delta, x) := -\nabla_d E(\psi(Y_\delta, \varphi_d(\delta, x)) \mid D = d, X = x)$ has full rank $\lambda$.

**Proof of Theorem 6.2:** Given B.1 and C.1-C.3, Theorem A.1 applies to ensure that given $\epsilon > 0$, if $\sigma_{S_d,D}(\delta, x) = \|1 - \zeta_d(\delta, x, \cdot)\|_2 < \eta_\epsilon$ for $\eta_\epsilon$ sufficiently small, then $\varphi_d(\delta, x)$ lies in an open $\epsilon$–neighborhood of $a = \varphi_d(\delta, x)$. Assumptions B.1, C.1, C.2, and C.4 then justify the application of Taylor’s theorem and the solution for $b_d(\delta, x)$ as in the text.

To obtain the bound, we use the fact that if $A, B$ are positive semi-definite $m \times m$ matrices, then $0 \leq tr(AB) \leq (trA)(trB)$ (e.g., Abadir and Magnus, 2005, p.329). Letting $\sigma_{\varepsilon S}(\delta, x) := E(\varepsilon S_{d,D} \mid D = \delta, X = x)$ and suppressing $(\delta, x)$, we have

$$b_d J_d' J_d b_d = \sigma_{\varepsilon S} J_d [J_d' J_d]^{-1} J_d' \sigma_{\varepsilon S} \sigma_{\varepsilon S} = \text{tr}(J_d [J_d' J_d]^{-1} J_d' \sigma_{\varepsilon S} \sigma_{\varepsilon S})$$

$$\leq \text{tr}(J_d [J_d' J_d]^{-1} J_d' \sigma_{\varepsilon S} \sigma_{\varepsilon S})$$

$$= \lambda \sigma_{\varepsilon S} \sigma_{\varepsilon S}.$$

Applying elementary inequalities and Cauchy-Schwarz to the last expression gives

$$[b_d(\delta, x)' J_d(\delta, x)' J_d(\delta, x) b_d(\delta, x)]^{1/2} \leq \lambda^{1/2} \sigma_{\varepsilon S}(\delta, x) \sigma_{S_d,D}(\delta, x).$$

The next two assumptions ensure the existence of derivatives and the interchange of derivative and integral for Theorem 6.3. For brevity, these conditions are not fully primitive. Sufficient primitive conditions can be readily given, but are rather lengthy. Again, $(d, x)$ is taken as given.

**Assumption C.5** For all $\epsilon$ such that $d + \epsilon \in \mathcal{C}$: (i) $E[\psi(r(d + \epsilon, Z, U), \varphi_d(d + \epsilon, X)) \mid D = d, X = x] < \infty$; and (ii) $\nabla_r \psi(r(d + \epsilon, z, u), \varphi_d(d + \epsilon, x)))$ exists for almost every $u$ in $\mathcal{U}_{d,x}$. (iii) There is a random variable $\Delta_{d,x}(U)$ with $E(\Delta_{d,x}(U) \mid D = d, X = x) < \infty$ such that for all $\epsilon$ such that $d + \epsilon \in \mathcal{C}$, $|\nabla_r \psi(r(d + \epsilon, z, u), \varphi_d(d + \epsilon, x))| \leq \Delta_{d,x}(u)$ for almost every $u$ in $\mathcal{U}_{d,x}$.

**Assumption C.6** For all $\epsilon$ such that $d + \epsilon \in \mathcal{C}$: (i) $\int_{\mathcal{U}_{d,x}} \psi(r(d + \epsilon, z, u), \alpha(d + \epsilon, x)) g(u \mid d + \epsilon, x) \, dv(u \mid d, x) < \infty$; and (ii) $\nabla_r g(u \mid d + \epsilon, x)$ exists for almost every $u$ in $\mathcal{U}_{d,x}$. (iii) There is a measurable $\overline{\Delta}_{d,x}(\cdot)$ with $\int_{\mathcal{U}_{d,x}} \overline{\Delta}_{d,x}(u) \, dv(u \mid d, x) < \infty$ such that for all $\epsilon$ such that $d + \epsilon \in \mathcal{C}$, $|\nabla_r \psi(r(d + \epsilon, z, u), \varphi_d(d + \epsilon, x)) g(u \mid d + \epsilon, x)| \leq \overline{\Delta}_{d,x}(u)$ for almost every $u$ in $\mathcal{U}_{d,x}$.
Proof of Theorem 6.3: Differentiating $E(\psi(Y_\delta, \varphi_d(\delta, x)) \mid D = d, X = x) = 0$ gives

$$0 = \nabla_\delta \int_{U_{d,x}} \psi(r(\delta, z, u), \varphi_d(\delta, x)) \, dG(u \mid d, x)$$

$$= \int_{U_{d,x}} \nabla_\delta \left[ \psi(r(\delta, z, u), \varphi_d(\delta, x)) \right] \, dG(u \mid d, x)$$

$$= \int_{U_{d,x}} \nabla_\delta r(\delta, z, u) \nabla_y \psi(r(\delta, z, u), \varphi_d(\delta, x)) \, dG(u \mid d, x)$$

$$+ \int_{U_{d,x}} \nabla_\delta \varphi_d(\delta, x) \nabla_a \psi(r(\delta, z, u), \varphi_d(\delta, x)) \, dG(u \mid d, x).$$

Solving and evaluating at $\delta = d$, we have

$$\int_{U_{d,x}} \nabla_d r(d, z, u) \nabla_y \psi(r(d, z, u), \alpha(d, x)) \, dG(u \mid d, x)$$

$$= -\nabla_\delta \varphi_d(d, x) \int_{U_{d,x}} \nabla_a \psi(r(d, z, u), \alpha(d, x)) \, dG(u \mid d, x). \tag{3}$$

Applying B.4, differentiating the equation defining $\alpha$, and evaluating at $\epsilon = 0$ gives

$$0 = \nabla_\epsilon \int_{U_{d,x}} \psi(r(d + \epsilon, z, u), \alpha(d + \epsilon, x)) \, d\nu(u \mid d, x)$$

$$= \int_{U_{d,x}} \nabla_\epsilon \left[ \psi(r(d + \epsilon, z, u), \alpha(d + \epsilon, x)) \right] \, d\nu(u \mid d, x)$$

$$= \int_{U_{d,x}} \nabla_\epsilon r(d + \epsilon, z, u) \nabla_y \psi(r(d + \epsilon, z, u), \alpha(d + \epsilon, x)) \, d\nu(u \mid d, x)$$

$$+ \int_{U_{d,x}} \nabla_\epsilon g(u \mid d + \epsilon, x) \, d\nu(u \mid d, x)$$

$$= \int_{U_{d,x}} \nabla_d r(d, z, u) \nabla_y \psi(r(d, z, u), \alpha(d, x)) \, dG(u \mid d, x)$$

$$+ \nabla_d \alpha(d, x) \int_{U_{d,x}} \nabla_a \psi(r(d, z, u), \alpha(d, x)) \, dG(u \mid d, x)$$

$$+ \int_{U_{d,x}} \nabla_d \ln g(u \mid d, x) \, \psi(r(d, z, u), \alpha(d, x)) \, dG(u \mid d, x).$$

Solving, substituting (3), and letting $\tilde{J}_d(x) = -\int_{U_{d,x}} \nabla'_a \psi(r(d, z, u), \alpha(d, x)) \, dG(u \mid d, x)$ gives

$$\tilde{J}_d(x) \nabla'_d \alpha(d, x) = \tilde{J}_d(x) \nabla'_d \varphi_d(d, x) + \int_{U_{d,x}} \psi(r(d, z, u), \alpha(d, x)) \, dG(u \mid d, x).$$

With $\tilde{K}_d(d) := [\tilde{J}_d(d)' \tilde{J}_d(d)]^{-1} \tilde{J}_d(d)'$, we obtain

$$\nabla_d \alpha(d, x) = \nabla'_d \varphi_d(d, x) + \tilde{K}_d(x) \int_{U_{d,x}} \psi(r(d, z, u), \alpha(d, x)) \, dG(u \mid d, x).$$
It follows that \( \tilde{b}_d(x) = \nabla_d b(d, x) \) has the representation

\[
\tilde{b}_d(x) = \tilde{K}_d(x) \int_{d,x} \psi(r(d, z, u), \alpha(d, x)) \nabla'_d \ln g(u \mid d, x) \, dG(u \mid d, x)
\]

\[
= \tilde{K}_d(x) \, E(\varepsilon \mid D = d, X = x).
\]

The bound follows by the same argument as in Theorem 6.2.

We do not prove Proposition 8.1, as this is an immediate consequence of Proposition 8.2.

**Proof of Proposition 8.2:** Let \( \mathcal{D}^* \) be such that \( U = f(\{Y_d, d \in \mathcal{D}^*\}, X) \). By assumption, \( \{Y_d, d \in \mathcal{D}^*\} \perp D \mid X \), so by lemmas 4.1 and 4.2(i) of D79, \( \{Y_d, d \in \mathcal{D}^*\}, X \perp D \mid X \). As \( U = f(\{Y_d, d \in \mathcal{D}^*\}, X) \), it follows from lemma 4.2(i) of D79 that \( U \perp D \mid X \).

**Acknowledgments:** The authors are indebted to the editor and two anonymous referees for their constructive suggestions. We also thank Patrick Bajari, Clive W.J. Granger, Jinyong Hahn, Lars Hansen, James Heckman, Kei Hirano, Stefan Hoderlein, Guido Imbens, Kevin Hoover, Meng Huang, Charles Manski, Massimiliano Marinucci, Robert Marshall, Rosa Matzkin, Whitney Newey, Judea Pearl, Andres Santos, Max Stinchcombe, Chris Taber, and Rosalind Wu for helpful discussions and comments. Any errors remain the authors’ responsibility.

**References**


Econometrica, 11, 1-12.


References


