Contents

3
Classical causal effects strategies

3.1 Causal effects and treatment effects

When evaluating accounting choices, we're deeply interested in their welfare effects. Does choice A make everyone better off or worse off compared with choice B? Or, does one choice make some better off and the other choice make others better off such that self-selection is a Pareto improvement? These are difficult questions and their resolution is invariably controversial. The root of the inference or modeling problem can be traced back to omitted, correlated regressor variables, as discussed in the above simpler settings.

Causal effects may involve choices with which we have experience in familiar environments or in new environments. Or, we may be interested in welfare effects associated with choices with which we have no experience in familiar or new environments. In the former case, where we have history on our side, we might pose treatment effect questions and employ historical data to help make an assessment. Treatment effects ask whether an individual's welfare is greater with treatment than without treatment. That is, other things are held constant and we attempt to explore the impact of treatment on welfare.

Treatment effects are less demanding than causal effects of unexplored choices in new environments. Nonetheless, treatment effect analysis poses serious challenges. The endogenous nature of choice often makes it difficult to hold other things constant. Observable outcome typically is an incomplete and ill-timed measure of welfare. While we're interested in the individual's expected utility, we usually observe ex post outcomes. Ex post versus ex ante considerations may or may not be easily surmounted. Outcomes may represent gross gains rather than ex ante differences in utility. Gross gains may be related to net benefits if costs are well understood but individual specific features (for example, nonpecuniary considerations) may be particularly elusive. One of the most severe challenges is we typically observe data for an individual only with treatment or without treatment but not for both. This implies that we cannot directly assess an individual's treatment effect. However, homogeneity conditions may allow inference based on population-level treatment effect parameters (for example, mean or average treatment effects).

3.2 A simple treatment effect example

The above *ANCOVA* example illustrates a simple treatment effect analysis if, for instance, counterfactuals have the same probability distribution as those observed. Counterfactuals are conditions not observed. To fix ideas, let Y_1 denote outcome with treatment and Y_0 outcome without treatment. Then the treatment effect is $Y_1 - Y_0$. However, we observe $(Y_1 \mid D = 1)$ and $(Y_0 \mid D = 0)$ but don't observe the counterfactuals, $(Y_1 \mid D = 0)$ and $(Y_0 \mid D = 1)$. We would like to compare outcome with treatment to outcome without treatment for individuals who chose treatment (treatment effect on the treated — TT) and for individuals who chose no treatment (treatment effect on the untreated — TUT). Both treatment effects compare factual with counterfactual outcomes

$$\begin{aligned} TT &= (Y_1 \mid D = 1) - (Y_0 \mid D = 1) \\ &= (Y_1 - Y_0 \mid D = 1) \end{aligned}$$

and

$$\begin{aligned} TUT &= (Y_1 \mid D = 0) - (Y_0 \mid D = 0) \\ &= (Y_1 - Y_0 \mid D = 0) \end{aligned}$$

Suppose we have the following DGP (factual and counterfactual)

$$\begin{aligned} E\left[Y_1 \mid D = 1, X = -1\right] &= E\left[Y_1 \mid D = 0, X = -1\right] \\ &= E\left[Y_1 \mid X = -1\right] = 9 \end{aligned}$$

$$\begin{aligned} E\left[Y_1 \mid D = 1, X = 0\right] &= E\left[Y_1 \mid D = 0, X = 0\right] \\ &= E\left[Y_1 \mid X = 0\right] = 10 \end{aligned}$$

$$\begin{aligned} E\left[Y_1 \mid D = 1, X = 1\right] &= E\left[Y_1 \mid D = 0, X = 1\right] \\ &= E\left[Y_1 \mid X = 1\right] = 11 \end{aligned}$$

$$\begin{aligned} E\left[Y_0 \mid D=1, X=-1\right] &= E\left[Y_0 \mid D=0, X=-1\right] \\ &= E\left[Y_0 \mid X=-1\right]=4 \end{aligned}$$

$$\begin{aligned} E\left[Y_0 \mid D=1, X=0\right] &= E\left[Y_0 \mid D=0, X=0\right] \\ &= E\left[Y_0 \mid X=0\right]=5 \end{aligned}$$

$$\begin{aligned} E\left[Y_0 \mid D=1, X=1\right] &= E\left[Y_0 \mid D=0, X=1\right] \\ &= E\left[Y_0 \mid X=1\right]=6 \end{aligned}$$

Treatment is said to be ignorable or selection is on observables as the regressors are sufficiently informative to make treatment, D, conditionally uninformative of outcome with treatment, Y_1, and outcome without treatment, Y_0. Then, the conditional average treatment effects on the treated $(ATT\left(X\right))$ and on the untreated $(ATUT\left(X\right))$ are

$$\begin{aligned} ATT\left(X=-1\right) &= E\left[Y_1-Y_0 \mid D=1, X=-1\right] \\ &= 9-4=5 \end{aligned}$$

$$\begin{aligned} ATT\left(X=0\right) &= E\left[Y_1-Y_0 \mid D=1, X=0\right] \\ &= 10-5=5 \end{aligned}$$

$$\begin{aligned} ATT\left(X=1\right) &= E\left[Y_1-Y_0 \mid D=1, X=1\right] \\ &= 11-6=5 \end{aligned}$$

$$\begin{aligned} ATUT\left(X=-1\right) &= E\left[Y_1-Y_0 \mid D=0, X=-1\right] \\ &= 9-4=5 \end{aligned}$$

$$\begin{aligned} ATUT\left(X=0\right) &= E\left[Y_1-Y_0 \mid D=0, X=0\right] \\ &= 10-5=5 \end{aligned}$$

$$\begin{aligned} ATUT\left(X=1\right) &= E\left[Y_1-Y_0 \mid D=0, X=1\right] \\ &= 11-6=5 \end{aligned}$$

If outcome represents net benefit, then, conditional on X, everyone is better off with treatment than without treatment. Since this is true for all levels of X, it is not surprising that, on applying iterated expectations, the unconditional average treatment effects on the treated (ATT) and on the untreated ($ATUT$) indicate an average (over all common X) net benefit as well.[1]

$$\begin{aligned} ATT &= E_X\left[E\left[Y_1 - Y_0 \mid D = 1, X\right]\right] \\ &= E\left[Y_1 - Y_0 \mid D = 1\right] = 5 \end{aligned}$$

and

$$\begin{aligned} ATUT &= E_X\left[E\left[Y_1 - Y_0 \mid D = 0, X\right]\right] \\ &= E\left[Y_1 - Y_0 \mid D = 0\right] = 5 \end{aligned}$$

Of course, this degree of homogeneity implies the average treatment effect is

$$\begin{aligned} ATE &= \Pr\left(D = 1\right) ATT + \left(1 - \Pr\left(D = 1\right)\right) ATUT \\ &= \Pr\left(D = 1\right) E\left[Y_1 - Y_0 \mid D = 1\right] + \Pr\left(D = 0\right) E\left[Y_1 - Y_0 \mid D = 0\right] \\ &= E\left[Y_1 - Y_0\right] = 5 \end{aligned}$$

3.3 Treatment effects with limited common support

Unfortunately, the above DGP, where outcome reflects welfare, outcome is homogeneous, and common X support, is rarely encountered. Rather, it's typical to encounter some heterogeneity in outcome and limited common support.[2] To illustrate the implications of limited common support, suppose we have the following data (where relative population frequencies are reflected by their sample frequencies).

Y	Y_1	Y_0	D	X
4	11	4	0	0
6	12	6	0	-1
5	13	5	0	-1
4	11	4	0	0
11	11	4	1	0
11	11	4	1	0
9	9	3	1	1
10	10	2	1	1

[1] Common support for X is important as our inferences stem from evidence we have rather than evidence we don't have in hand.

[2] Further, often outcome measures gross benefits (and perhaps incompletely) rather than net benefits so that welfare implications require knowledge of costs with and without treatment.

We don't observe the counterfactuals: $(Y_1 \mid D=0)$ or $(Y_0 \mid D=1)$, but the key to identifying any average treatment effect is

$$E\left[Y_1 \mid X=x, D=1\right] = E\left[Y_1 \mid X=x, D=0\right]$$

and

$$E\left[Y_0 \mid X=x, D=1\right] = E\left[Y_0 \mid X=x, D=0\right]$$

That is, the pivotal condition is outcome mean conditional independence of treatment, D. For the only commonly observed value, $x=0$

$$E\left[Y_1 \mid X=0, D=1\right] = E\left[Y_1 \mid X=0, D=0\right] = 11$$

and

$$E\left[Y_0 \mid X=0, D=1\right] = E\left[Y_0 \mid X=0, D=0\right] = 4$$

conditional mean independence is satisfied. Hence, the only evidence-based assessment of the treatment effect is for $X=0$, and

$$ATE\left(X=0\right) = E\left[Y_1 - Y_0 \mid X=0\right] = 11 - 4 = 7$$

Further, this conditional average treatment effect is homogeneous.

$$ATT\left(X=0\right) = ATUT\left(X=0\right) = ATE\left(X=0\right) = 7$$

where

$$ATT\left(X=0\right) = E\left[Y_1 - Y_0 \mid X=0, D=1\right] = 11 - 4 = 7$$

and

$$ATUT\left(X=0\right) = E\left[Y_1 - Y_0 \mid X=0, D=0\right] = 11 - 4 = 7$$

While this conditional average treatment effect is, in principle, only nonparametrically identified, by good fortune, *ANCOVA* effectively estimates both the conditional (on $X=0$) and unconditional average treatment effect via the coefficient on D.[3]

$$E\left[Y \mid D, X\right] = 4 + 7D - 1.5X$$

where the observables are

$$Y = DY_1 + \left(1-D\right)Y_0$$

Further, the conditional average treatment effect also equals the unconditional average. Since the unconditional average treatment effect is unidentified by observable data, both of these results are merely fortuitous. That

[3] More generally, we include an interaction term, $\left(D \times \left(X - \overline{X}\right)\right)$, but it's coefficient is zero for this *DGP*.

is, the only conclusion we can draw based on the evidence is for the average treatment effect conditional on $X = 0$. If there is a local interval of common X support, this is sometimes called a local average treatment effect.

To clarify this common support issue, suppose we perturb only the counterfactual outcomes with treatment as follows.

Y	Y_1	Y_0	D	X
4	11	4	0	0
6	2	6	0	-1
5	1	5	0	-1
4	11	4	0	0
11	11	4	1	0
11	11	4	1	0
9	9	3	1	1
10	10	2	1	1

Now, the unconditional average treatment effect is $8.25 - 4 = 4.25$, the unconditional average treatment effect on the treated is unperturbed from above, $10.25 - 3.25 = 7$, and the unconditional average treatment effect on the untreated is $6.25 - 4.75 = 1.5$. Hence, outcome is heterogeneous and none of these unconditional average treatment effects are identified by the data. As above, the only treatment effect identifiable from the data is the conditional average treatment effect for $X = 0$, which continues to be $ATE\,(X = 0) = 11 - 4 = 7$. Attempting to extrapolate from the evidence to unconditional average treatment effects is not only a stab in the dark, it is misleading.

Next, we pursue a variation on treatment effects based on discontinuity induced by assignment variable(s). This is referred to as regression discontinuity design and can make identification of the associated treatment effects relatively straightforward.

3.4 Regression discontinuity designs

Suppose we have continuous assignment variable(s), Z, for which treatment is selected or assigned above a cutoff, z_0, and no treatment is selected or assigned below the cutoff. Then, the effect of treatment on potential outcomes with and without treatment can be identified in the region of the cutoff by simple mean differences (it's said to be nonparametrically identified) as any discontinuity is attributable to treatment. For such data generating processes (DGP), regression discontinuity (RD) designs identify a marginal treatment effect in the vicinity of the cutoff and, perhaps, a weighted average of heterogeneous treatment effects.[4] For a sharp RD

[4] See Lee and Lemieux [2010], p. 298 and a brief subsection under heterogeneous treatment effects.

design, no potential outcomes with treatment are observed below the cutoff and no potential outcomes without treatment are observed above the cutoff.

A so-called fuzzy RD design involves a discontinuity at the cutoff as with the sharp design, however the researcher doesn't observe all determinants of treatment. Consequently, treatment is stochastic rather than deterministic. The key feature of both RD designs is outcome discontinuity as a function of the assignment variables at the cutoff. The following describes observed outcome, Y_i, for both sharp and fuzzy RD designs.

$$\begin{aligned} Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\ &= \alpha_i + D_i \tau_i \end{aligned}$$

where $\alpha_i = Y_{0i}$ outcome without treatment, Y_{1i} is outcome with treatment, $D_i = 1$ for treated and zero otherwise, and $\tau_i = Y_{1i} - Y_{0i}$. Our discussion focuses primarily on sharp RD designs for simplicity and clarity. The sharp RD design identifies treatment effect

$$\tau = \frac{Y^+ - Y^-}{D^+ - D^-} = Y^+ - Y^-$$

since $D^+ - D^- = 1$ where

$$Y^+ = \lim_{z \overset{+}{\to} z_0} E[Y_i \mid Z_i = z]$$

$$D^+ = \lim_{z \overset{+}{\to} z_0} E[D_i \mid Z_i = z]$$

the limit approaches the cutoff from above,

$$Y^- = \lim_{z \overset{-}{\to} z_0} E[Y_i \mid Z_i = z]$$

and

$$D^- = \lim_{z \overset{-}{\to} z_0} E[D_i \mid Z_i = z]$$

the limit approaches the cutoff from below. Fuzzy RD designs are briefly discussed at the end of the homogeneous treatment effect section where fuzzy RD treatment effects are identified as

$$\tau = \tfrac{\tau_r}{\delta} = \tfrac{Y^+ - Y^-}{D^+ - D^-}, \quad \text{for } D^+ \neq D^-$$

the intent to treat effect is τ_r, and the propensity for treatment is δ (see Hahn, Todd, and van der Klaauw [2001] and Lee and Lemieux [2010]).

3.4.1 Homogeneous treatment effects

For homogeneous treatment effects, the key identification condition is outcome without treatment, Y_0, is smooth in the neighborhood of the cutoff z_0. To appreciate this identification strategy, explored as a thought experiment, recognize outcome with treatment, Y_1, is observed on one side of the cutoff (say, after the trigger) and Y_0 is observed before the trigger (cutoff). The treatment effect at the cutoff is $(Y_1 - Y_0 \mid Z = z_0)$ but this is not observed. However, we do observe $(Y_1 \mid Z = z_0 + \varepsilon)$ and $(Y_0 \mid Z = z_0 - \varepsilon)$. Hence, if we believe the *DGP* satisfies $(Y_0 \mid Z = z_0 - \varepsilon) \approx (Y_0 \mid Z = z_0 + \varepsilon)$; in other words, if there is a smooth transition in Y_0 from $z_0 - \varepsilon$ (observed) to $z_0 + \varepsilon$ (unobserved) then the homogeneous treatment effect is identified around the cutoff and can be estimated from observed data.

We consider two homogeneous treatment effect examples: one illustrating a *DGP* well-suited to *RD* design, and a second *DGP* where continuity of Y_0 and identification fails (or an alternative interpretation is individuals are able to "precisely manipulate" the assignment variable, Lee and Lemieux [2010]).

Example 1 *Suppose we have the following sharp RD representative data where $D = 1$ refers to treated and $D = 0$ refers to untreated.*

Y_0	Y_1	$Y_1 - Y_0$	Y	D	Z
14	15	1	15	1	5
13	14	1	14	1	4
11.5	12.5	1	12.5	1	3.501
11.5	12.5	1	11.5	0	3.499
10	11	1	10	0	3
9	10	1	9	0	2

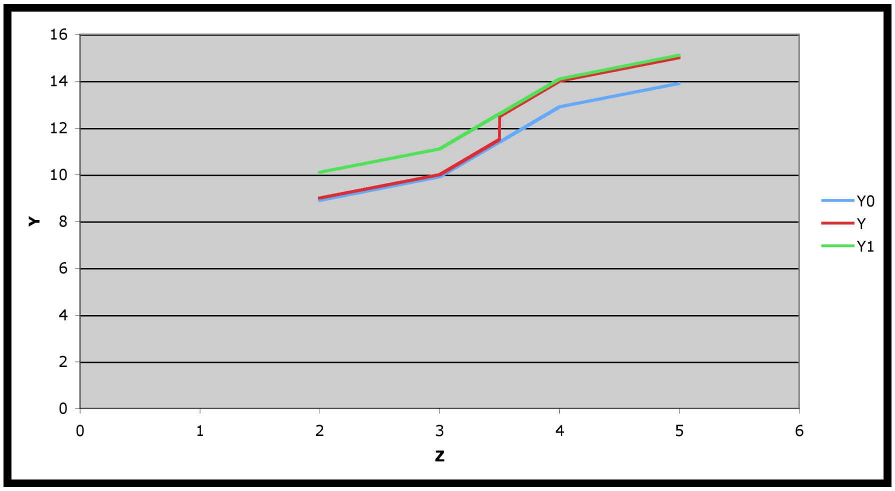
As is evident from the treatment effect column, $Y_1 - Y_0$, the treatment effect is homogeneous (for instance, treatment on treated and treatment on untreated are both equal to one). The marginal (and average, as all treatment effects are equal in this case) treatment effect at the cutoff identified by an RD design is

$$\begin{aligned} TE &\equiv \tau = \frac{Y^{+} - Y^{-}}{D^{+} - D^{-}} \\ &= E\left[Y_1 - Y_0 \mid Z \approx z_0\right] \\ &= \frac{12.5 - 11.5}{1 - 0} = 1 \end{aligned}$$

where ε defines a narrow region around z_0. Since $Y = DY_1 + (1 - D) Y_0$ is the only potential outcome that is observable, we can estimate the above treatment effect by

$$\begin{aligned} estTE &\equiv \widehat{\tau} = E\left[Y \mid Z = 3.501\right] - E\left[Y \mid Z = 3.449\right] \\ &= 12.5 - 11.5 = 1 \end{aligned}$$

This corresponds to the homogeneous treatment effect defined by comparison of observable outcomes with counterfactuals (unobservable potential outcomes). The data are nonlinear as illustrated in the graph below. The graph depicts observed outcome, Y, as well as potential outcomes with treatment, Y_1, and potential outcomes without treatment, Y_0, as a function of the assignment variable, Z. In principle, nonlinearity creates no problem for the RD design (we emphasize this to dispel nonlinearity as an explanation of the problem illustrated in example 2).[5]



Example 2 *On the other hand, suppose the representative data are slightly perturbed.*

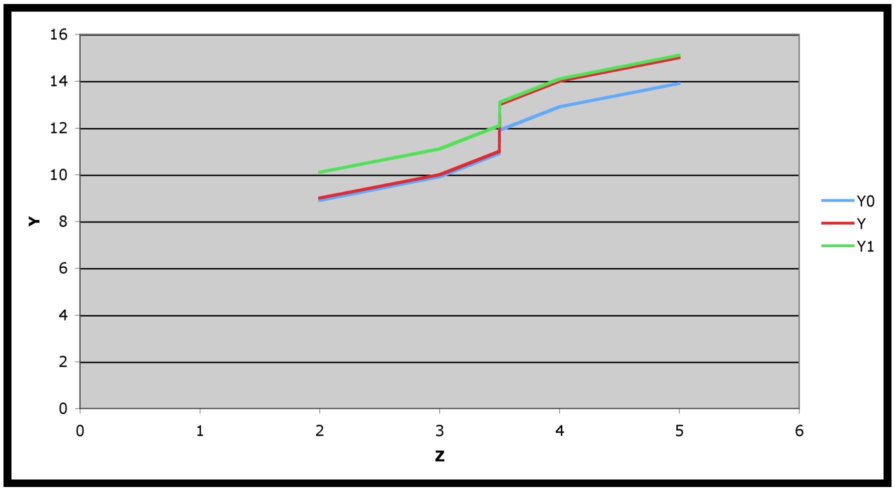
Y_0	Y_1	$Y_1 - Y_0$	Y	D	Z
14	15	1	15	1	5
13	14	1	14	1	4
12	13	1	13	1	3.501
11	12	1	11	0	3.499
10	11	1	10	0	3
9	10	1	9	0	2

There is a substantive jump in potential outcomes in the vicinity of the cutoff. Even though treatment effects continue to be homogeneous, RD design does not identify the treatment effect. Rather, the effect estimated by RD produces

$$\begin{aligned} estTE &= E[Y \mid Z = 3.501] - E[Y \mid Z = 3.449] \\ &= 13 - 11 = 2 \neq \tau = 1 \end{aligned}$$

[5] Potentially unknown functional form is why we say the RD treatment effect is nonparametrically identified. That is, nonparametric regression allows us to estimate the treatment effect in the neighborhood of the cutoff point by point.

This doesn't correspond to the homogeneous treatment effect defined by comparison of observable outcomes with counterfactuals. The graph below illustrates the discontinuity of Y_0 in Z around z_0; again, outcomes are nonlinear in the assignment variable but that is not the source of the problem.



Fuzzy RD design example

Next, we briefly illustrate a fuzzy *RD* design with homogeneous treatment effects.

Example 3 *Suppose we have the following fuzzy RD representative data.*

Y_0	Y_1	$Y_1 - Y_0$	Y	D	Z
14	15	1	15	1	5
13	14	1	14	1	4
11.5	12.5	1	12.5	1	3.501
11.5	12.5	1	11.5	0	3.501
11.5	12.5	1	11.5	0	3.499
10	11	1	10	0	3
9	10	1	9	0	2

As is evident from the treatment effect column, $Y_1 - Y_0$, the treatment effect is homogeneous. The fuzzy treatment effect identified by a fuzzy RD design is

$$\begin{aligned} TE &\equiv \tau = \frac{\tau_r}{\delta} \\ &= \frac{Y^+ - Y^-}{D^+ - D^-} \\ &= E\left[Y_1 - Y_0 \mid Z \approx z_0\right] \\ &= \frac{0.5}{0.5} = 1 \end{aligned}$$

Since $Y = DY_1 + (1 - D) Y_0$ *is the only potential outcome that is observable, we can estimate the above treatment effect by*

$$\begin{aligned} estTE &\equiv \widehat{\tau} = \frac{E[Y \mid Z = 3.501] - E[Y \mid Z = 3.449]}{E[D \mid Z = 3.501] - E[D \mid Z = 3.449]} \\ &= \frac{12 - 11.5}{0.5 - 0.0} = 1 \end{aligned}$$

An "intent to treat" effect can be identified from the reduced form (Lee and Lemieux [2010], p. 328). The reduced form follows from substituting the selection equation

$$D^* = \gamma + \delta T + g(Z - z_0) + \nu$$

into the outcome equation

$$Y = \alpha + \tau D^* + f(Z - z_0) + \varepsilon$$

to yield

$$Y = \alpha_r + \tau_r T + f_r(Z - z_0) + \varepsilon_r$$

where $T = 1[Z \geq z_0]$*. The coefficient on* T*,* $\tau_r = \tau\delta = 0.5$*, is the intent to treat effect (a rescaling of the treatment effect,* τ*, by the propensity for treatment around the cutoff,* δ*).*

3.4.2 Heterogeneous treatment effects

Hahn, Todd, and van der Klaauw [2001] propose two alternative strategies for identification of treatment effects where potential outcomes are heterogeneous. The first strategy adopts an ignorable treatment approach similar to that invoked by matching strategies. The second strategy may be more amenable to settings of self-selection as it identifies a discrete marginal treatment effect (a local average treatment effect or $LATE$, for short) along the lines of Imbens and Angrist [1994].

Conditions on the DGP for the ignorable treatment strategy are
(a) continuity of Y_0 in the neighborhood of the cutoff,
(b) continuity of the treatment effect, τ, in the neighborhood of the cutoff (again as a function of the assignment variables, Z), and
(c) conditional independence of treatment D_i and the treatment effect τ_i in the neighborhood of the cutoff, z_0.

The conditional independence condition maintains that individuals do not select treatment in anticipation of gains from treatment which may run counter to self-selection. The $LATE$ strategy replaces conditions (b) and (c) above with
(b') $(\tau_i, D_i(Z))$ are jointly independent of Z_i near z_0, and
(c') $D_i(z_0 + e) \geq D_i(z_0 - e)$ for all $0 < e < \varepsilon$. This implies one way flows into or away from treatment by varying the assignment variable in the

neighborhood of the cutoff. Imbens and Angrist [1994] refer to this condition as monotonicity, while Heckman and Vytlacil [2005] refer to this as uniformity. This condition is always satisfied, by definition, for a sharp *RD* design but implies *LATE* is defined only for a subpopulation of "compliers" in a fuzzy *RD* design.

Next, we consider two heterogeneous treatment effect examples: one illustrating a *DGP* well-suited to sharp *RD* design, and a second *DGP* where continuity/smoothness of the treatment effect fails. The latter suggests how an *RD* design might fail to identify the (marginal) effect of treatment.

Example 4 *Suppose we have the following representative heterogeneous data.*

Y_0	Y_1	$Y_1 - Y_0$	Y	D	Z
10	15	5	15	1	5
11	14	3	14	1	4
11.5	12.5	1	12.5	1	3.501
11.5	12.5	1	11.5	0	3.499
13	12	-1	13	0	3
14	11	-3	14	0	2

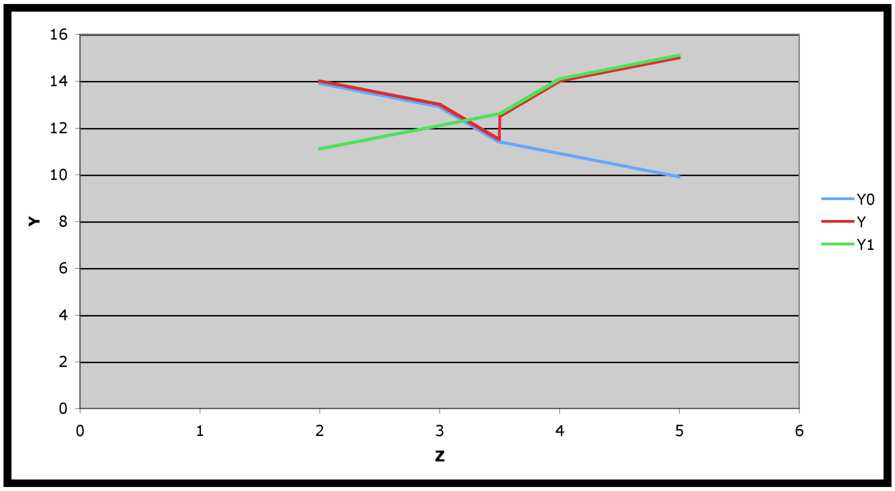
As is evident from the treatment effect column, $Y_1 - Y_0$, the treatment effect is heterogeneous (for instance, treatment on treated and treatment on untreated are unequal). The (marginal) treatment effect identified by an RD design is

$$\begin{aligned} TE &\equiv \tau = Y^+ - Y^- \\ &= E\left[Y_1 - Y_0 \mid Z \approx z_0\right] = 1 \end{aligned}$$

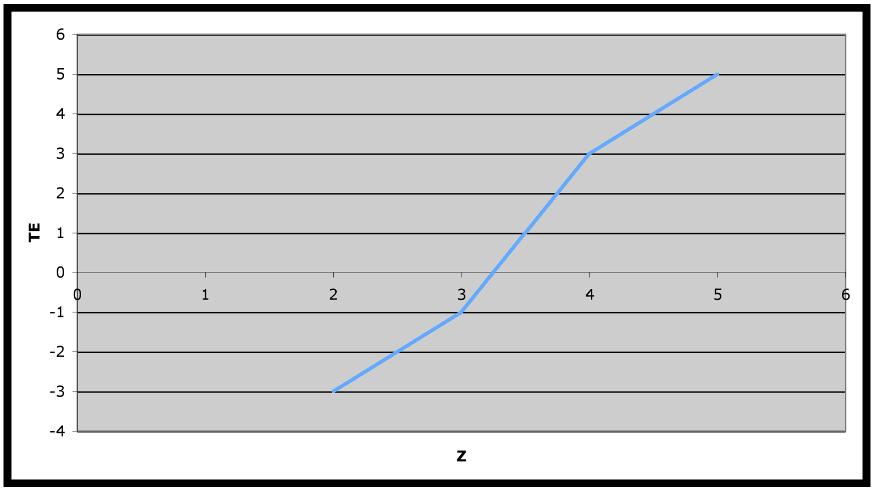
where ε defines a narrow region of z_0. Since Y is the only potential outcome that is observable, we can estimate the above treatment effect by nonparametric local linear regression (that is, point by point estimation to accommodate unknown functional form in the neighborhood of the cutoff, z_0).

$$\begin{aligned} estTE &\equiv \widehat{\tau} = E\left[Y \mid Z = 3.501\right] - E\left[Y \mid Z = 3.449\right] \\ &= 12.5 - 11.5 = 1 \end{aligned}$$

This corresponds to the marginal treatment effect (LATE) defined by comparison of observable outcomes with counterfactuals (unobservable potential outcomes) in the vicinity of z_0. The data are nonlinear as illustrated in the graph below. The graph depicts observed outcome, Y, as well as potential outcomes with treatment, Y_1, and potential outcomes without treatment, Y_0, as a function of the assignment variable, Z. While nonlinearity creates no insurmountable problem for the RD design, to accommodate unknown functional form consistent estimation might employ nonparametric local linear *regression.*



The above graph provides an idea of the potential outcomes. The key is smoothness of the (unobservable) treatment effect or the independence of the treatment effect and the assignment variables in the neighborhood of the cutoff. This is illustrated below for the DGP.



Treatment effects are clearly smooth in the assignment variable, Z, in the vicinity of the cutoff, $z_0 = 3.5$.

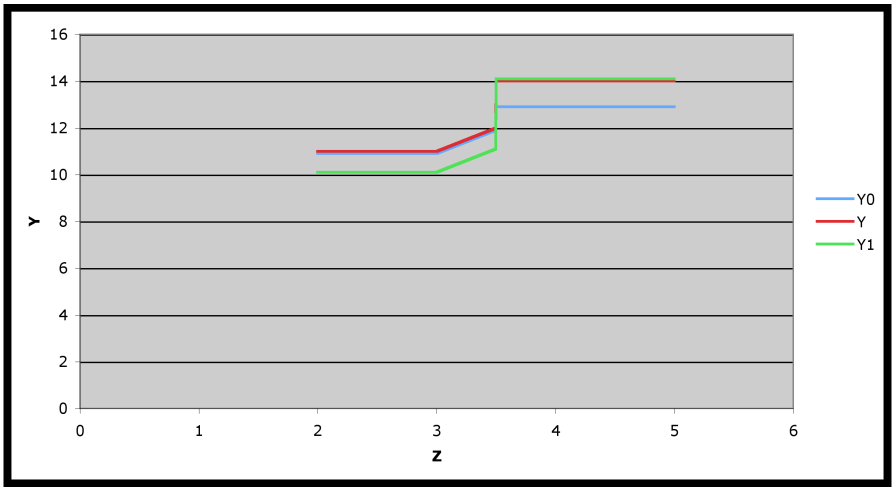
Example 5 *On the other hand, suppose the representative data are perturbed such that treatment on treated is positive and treatment on untreated* *is negative as depicted below.*

Y_0	Y_1	$Y_1 - Y_0$	Y	D	Z
13	14	1	14	1	5
13	14	1	14	1	4
13	14	1	14	1	3.501
12	11	-1	12	0	3.499
11	10	-1	11	0	3
11	10	-1	11	0	2

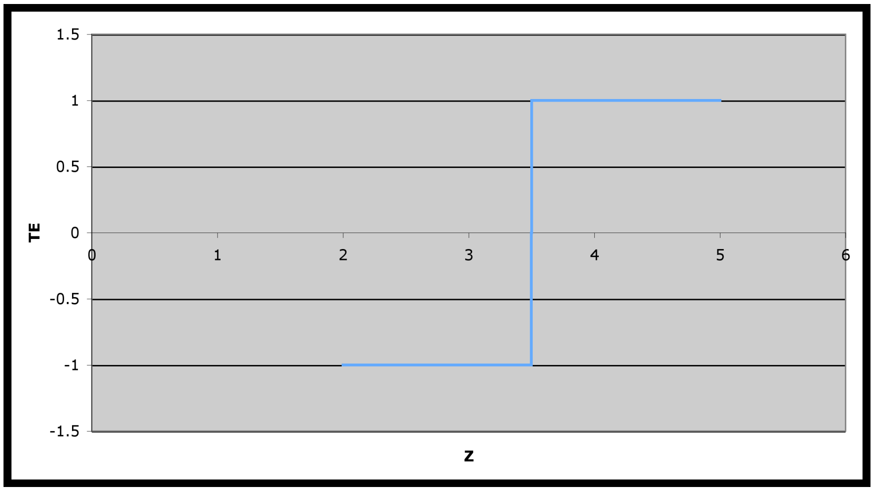
There is clearly a discontinuous jump in the treatment effect in the vicinity of the cutoff. While treatment effects are not as erratic as the previous example, the RD strategy does not identify the marginal treatment effect. Rather, the treatment effect estimated by the above strategy produces

$$\begin{aligned} estTE &\equiv \widehat{\tau} = E\left[Y \mid Z = 3.501\right] - E\left[Y \mid Z = 3.449\right] \\ &= 14 - 12 = 2 \neq \tau = E\left[Y_1 - Y_0 \mid Z \approx z_0\right] = 0 \end{aligned}$$

This doesn't correspond to the heterogeneous treatment effect defined by (point by point) comparison of observable outcomes with counterfactuals (in the neighborhood of the cutoff, the treatment effect average is zero). The graph below illustrates nonlinearity of outcomes but that is not the source of the problem.



The key picture depicts the relation between (unobservable) treatment effects and the assignment variable. This is presented below.



Clearly, the relationship between the treatment effect and the assignment variable is not smooth and the treatment effect is not independent of the assignment variables in the neighborhood of the cutoff. Consequently, the regression discontinuity design fails to identify the marginal treatment effect for compliers (individuals who select or are assigned treatment just above the cutoff and who select or are assigned no treatment just below the cutoff).

Weighted average of heterogeneous treatment effects

Lee and Lemieux [2010, p. 298] suggest the RD designs identify a weighted average of heterogeneous treatment effects. Define outcome with unrestricted heterogeneity

$$Y = D\tau(Z, U) + Z\delta_1 + U$$

and selection

$$D^* = Z\delta_2 + V$$

Now, Bayesian manipulation gives

$$\begin{aligned} & \lim_{\varepsilon \xrightarrow{+} 0} E[Y_i \mid Z_i = z_0 + \varepsilon] - \lim_{\varepsilon \xrightarrow{-} 0} E[Y_i \mid Z_i = z_0 - \varepsilon] \\ = & \sum_{z,u} \tau(z, u) \Pr(Z = z, U = u \mid D^* = z_0) \\ = & \sum_{z,u} \tau(z, u) \Pr(Z = z, U = u) \frac{f_{D^*|Z,U}(z_0 \mid Z = z, U = u)}{f_{D^*}(z_0)} \end{aligned}$$

Except for weighting by $\frac{f(z_0|Z=z,U=u)}{f(z_0)}$ this expression is the standard unconditional average treatment effect; hence, we have a weighted average of potentially heterogeneous treatment effects.

3.4.3 Estimation

Sharp and fuzzy *RD* design treatment effects are nonparametrically identified. Since local linear nonparametric regressions (*LLR*) are better behaved at the boundary and converge more quickly than standard kernel density nonparametric regressions, *LLR* is expected to have better small sample properties (Hahn, Todd, and van der Klaauw [2001]).[6] Y^+ is estimated by $\widehat{a^+}$ in

$$\left(\widehat{a^+}, \widehat{b^+}\right) \equiv \arg\min_{a,b} \sum_{i=1}^{n} \left(Y_i - a - b\left(Z_i - z_0\right)\right)^2 K\left(\frac{Z_i - z_0}{h}\right) 1\left(Z_i > z_0\right)$$

and Y^- is estimated by $\widehat{a^-}$ in

$$\left(\widehat{a^-}, \widehat{b^-}\right) \equiv \arg\min_{a,b} \sum_{i=1}^{n} \left(Y_i - a - b\left(Z_i - z_0\right)\right)^2 K\left(\frac{Z_i - z_0}{h}\right) 1\left(Z_i < z_0\right)$$

where $K(\cdot)$ is a kernel density and $h > 0$ is a suitable bandwidth. Hence, the treatement effect is estimated by $\widehat{\tau} = \widehat{a^+} - \widehat{a^-}$. Bootstrap inference is now straightforward.

3.5 Synthetic controls

Suppose one has panel data but very limited evidence on the treated and/or untreated subpopulations (for example, perhaps only one individual is treated). Then, standard matching strategies (general or propensity score) are likely of limited utility. On the other hand, perhaps one can construct a match for the treated based on a composite of the controls (untreated). This is an extension of difference-in-differences (*D-I-D*) in which unobservable factors (perhaps correlated with covariates, X_{it}) are allowed to vary cross-sectionally whereas such unobservables are a cross-sectional constant for the *D-I-D* strategy. This is commonly referred to as a factor model since the unobservables can be thought of as factor scores, λ_t, (which vary through time) weighted by factor loadings, μ_i, (which vary cross-sectionally) both of which are unobservable in a factor analysis.

$$Y_{it} = \alpha d_{it} + \beta^T X_{it} + \lambda_t \mu_i + \varepsilon_{it}$$

[6] *LLR* is an approximation. In other words, the flexibility of nonparametric regression does not imply that the unknown functional form is completely resolved — bias remains.

where Y_{it} is outcome for individual i during period t, β is an unknown vector of coefficients common across individuals, ε_{it} is unobservable (random) noise, d_{it} is an indicator of treatment (1) or nontreatment (0), and α is the treatment effect of interest.

Synthetic controls are constructed based on a convex combination (with weights $w_i \geq 0$, $\sum w_i = 1$) of the pre-intervention outcomes, $\widehat{Y}_i^k = \sum_{t=1}^{T_0} k_t Y_{it}$, and covariates, X_i, from the control subpopulation such that

$$\arg\min_W \left(H_1 - [H_2 \cdots H_n] W\right)^T V \left(H_1 - [H_2 \cdots H_n] W\right)$$

where $H_i = \left[\widehat{Y}_i^{k_1}, \widehat{Y}_i^{k_2}, \ldots \widehat{Y}_i^{k_m}, X_i^T\right]^T$, $i = 1$ denotes treated while $i = 2, \ldots, n$ denotes control, and V is a diagonal matrix controlling potentially different weights applied to the components of H_i. Abadie et al [2010, 2011] propose choosing V to minimize the expected mean square error

$$\arg\min_V \left(Y_1 - Y_0 W^* (V)\right)^T \left(Y_1 - Y_0 W^* (V)\right)$$

during the pre-intervention period where Y_1 refers to outcomes for the treated and Y_0 outcomes for the control group.

Next, we explore several variations via some highly stylized examples. First, we visit examples with constant factors such that both *D-I-D* and synthetic controls identify the treatment effect. Then, we visit settings with heterogeneous factors in which *D-I-D* fails but synthetic controls effectively identify the treatment effect. Finally, we consider settings where neither *D-I-D* nor synthetic variables effectively identify the treatment effect. The first failure is due to violation of conditional mean independence of potential outcome without treatment while the second failure is due to the treated individuals' H lying outside the convex hull of the controls. In the latter case, synthetic controls can produce greater selection bias than *D-I-D*.

3.5.1 Examples

Suppose there are three individuals who are untreated during the first two periods but the third individual adopts treatment in the third period. The data generating process (*DGP*) is

$$Y = I_1 + I_2 + I_3 + T + 2I_3 \times T + \lambda\mu$$

where Y is observed outcome, I_j denotes and indicator variable equal to 1 for individual j and 0 otherwise, and T is an indicator equal to one for intervention period three and zero otherwise. In the examples below, Y_0 represents potential outcome without treatment and only $\lambda\mu$ (and, of course, outcome) varies.

Example 6 (*D-I-D* setting) *Suppose the DGP is*

Y_0	Y	I_1	I_2	I_3	T	$d = I_3 \times T$	$\lambda\mu$
2	2	1	0	0	0	0	1
2	2	1	0	0	0	0	1
3	3	1	0	0	1	0	1
2	2	0	1	0	0	0	1
2	2	0	1	0	0	0	1
3	3	0	1	0	1	0	1
2	2	0	0	1	0	0	1
2	2	0	0	1	0	0	1
3	5	0	0	1	1	1	1

$\alpha = 5 - 3 = 2$. *D-I-D identifies*

$$E[Y \mid X] = 2I_1 + 2I_2 + 2I_3 + T + 2I_3 \times T$$

That is, even though the nonzero error term biases the coefficients on the individual mean effects, the treatment effect is effectively identified. Any weights V and W produce synthetic controls that effectively identify the treatment effect.

Example 7 (*D-I-D* setting with time variation) *Suppose the DGP is*

Y_0	Y	I_1	I_2	I_3	T	$d = I_3 \times T$	$\lambda\mu$
-2	-2	1	0	0	0	0	-3
-2	-2	1	0	0	0	0	-3
4	4	1	0	0	1	0	2
-2	-2	0	1	0	0	0	-3
-2	-2	0	1	0	0	0	-3
4	4	0	1	0	1	0	2
-2	-2	0	0	1	0	0	-3
-2	-2	0	0	1	0	0	-3
4	6	0	0	1	1	1	2

$\alpha = 6 - 4 = 2$. *D-I-D identifies*

$$E[Y \mid X] = -2I_1 - 2I_2 - 2I_3 + 6T + 2I_3 \times T$$

That is, even though the nonzero error term biases the coefficients on the individual and time mean effects, the treatment effect is effectively identified. Again, any weights V and W produce synthetic controls that effectively identify the treatment effect.

Example 8 (synthetic control setting $W^T = \begin{bmatrix} 0.1 & 0.9 \end{bmatrix}$) *Suppose the DGP is*

Y_0	Y	I_1	I_2	I_3	T	$d = I_3 \times T$	$\lambda\mu$
-2	-2	1	0	0	0	0	-3
-2	-2	1	0	0	0	0	-3
4	4	1	0	0	1	0	2
4	4	0	1	0	0	0	3
4	4	0	1	0	0	0	3
0	0	0	1	0	1	0	-2
3.4	3.4	0	0	1	0	0	2.4
3.4	3.4	0	0	1	0	0	2.4
0.4	2.4	0	0	1	1	1	-1.6

$\alpha = 2.4 - 0.4 = 2$. *D-I-D identifies*

$$E[Y \mid X] = -\frac{1}{3}I_1 + 2\frac{1}{3}I_2 + 3\frac{2}{5}I_3 + T - 2I_3 \times T$$

Cross-sectional variation in the unobservable factors, $\lambda\mu$, produces substantial bias in the D-I-D estimate of the treatment effect. That is, it is so underestimated the sign is reversed. On the other hand, since the test subject lies in a convex hull of the control subjects' predicted outcomes, the synthetic control strategy effectively identifies the treatment effect.

	V fixed	*V estimated*
$V\begin{pmatrix} I_1 \\ I_2 \\ \widehat{Y}_i^{t=1} \\ \widehat{Y}_i^{t=2} \end{pmatrix}$	$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.335 & 0 & 0 & 0 \\ 0 & 0.037 & 0 & 0 \\ 0 & 0 & 0.332 & 0 \\ 0 & 0 & 0 & 0.296 \end{bmatrix}$
W^T	$\begin{bmatrix} 0.1 & 0.9 \end{bmatrix}$	$\begin{bmatrix} 0.1 & 0.9 \end{bmatrix}$
$\alpha = Y_{33} - W^T \begin{bmatrix} Y_{13} \\ Y_{23} \end{bmatrix}$	2	2

where $\widehat{Y}_i^{t=1} = \sum_{t=1}^{2} k_t Y_{it} = Y_{i1}$ *and* $\widehat{Y}_i^{t=2} = \sum_{t=1}^{2} k_t Y_{it} = Y_{i2}$.[7]

[7] Identification is also effective with $\widehat{Y}_i^{t=1} = \sum_{t=1}^{2} k_t Y_{it} = \frac{1}{2}(Y_{i1} + Y_{i2})$ (or, for that matter, any convex combination of Y_{i1} and Y_{i2}) in place of both $\widehat{Y}_i^{t=1}$ and $\widehat{Y}_i^{t=2}$.

Example 9 (synthetic control setting $W^T = \left[\begin{array}{cc} 0.9 & 0.1 \end{array}\right]$) *Suppose the DGP is*

Y_0	Y	I_1	I_2	I_3	T	$d = I_3 \times T$	$\lambda\mu$
-2	-2	1	0	0	0	0	-3
-2	-2	1	0	0	0	0	-3
4	4	1	0	0	1	0	2
4	4	0	1	0	0	0	3
4	4	0	1	0	0	0	3
0	0	0	1	0	1	0	-2
-1.4	-1.4	0	0	1	0	0	-2.4
-1.4	-1.4	0	0	1	0	0	-2.4
3.6	5.6	0	0	1	1	1	-1.6

$\alpha = 5.6 - 3.6 = 2$. *D-I-D identifies*

$$E\left[Y \mid X\right] = -1\frac{1}{3}I_1 + 1\frac{1}{3}I_2 - 2\frac{2}{5}I_3 + T + 6I_3 \times T$$

Cross-sectional variation in the unobservable factors, $\lambda\mu$, produces substantial (overestimation) bias in the D-I-D estimate of the treatment effect. On the other hand, since the test subject lies in a convex hull of the control subjects' predicted outcomes, the synthetic control strategy effectively identifies the treatment effect.

	V fixed	*V estimated*
$V\left(\begin{array}{c} I_1 \\ I_2 \\ \widehat{Y}_i^{t=1} \\ \widehat{Y}_i^{t=2} \end{array}\right)$	$\left[\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{array}\right]$	$\left[\begin{array}{cccc} 0.037 & 0 & 0 & 0 \\ 0 & 0.335 & 0 & 0 \\ 0 & 0 & 0.332 & 0 \\ 0 & 0 & 0 & 0.296 \end{array}\right]$
W^T	$\left[\begin{array}{cc} 0.9 & 0.1 \end{array}\right]$	$\left[\begin{array}{cc} 0.9 & 0.1 \end{array}\right]$
$\alpha = Y_{33} - W^T \left[\begin{array}{c} Y_{13} \\ Y_{23} \end{array}\right]$	2	2

where $\widehat{Y}_i^{t=1} = \sum_{t=1}^{2} k_t Y_{it} = Y_{i1}$ *and* $\widehat{Y}_i^{t=2} = \sum_{t=1}^{2} k_t Y_{it} = Y_{i2}$.

Example 10 (both *d-i-d* and synthetic controls fail) *Suppose the DGP is almost identical to example 6 except that potential outcome without treat-*

ment is affected by the intervention (not conditionally mean independent).

Y_0	Y	I_1	I_2	I_3	T	$d = I_3 \times T$	$\lambda\mu$
2	2	1	0	0	0	0	1
2	2	1	0	0	0	0	1
3	3	1	0	0	1	0	1
2	2	0	1	0	0	0	1
2	2	0	1	0	0	0	1
3	3	0	1	0	1	0	1
2	2	0	0	1	0	0	1
2	2	0	0	1	0	0	1
5	5	0	0	1	1	1	1

$\alpha = 5 - 5 = 0$. *As before, D-I-D identifies*

$$E[Y \mid X] = 2I_1 + 2I_2 + 2I_3 + T + 2I_3 \times T$$

but the treatment effect is not 2 *but rather zero. Further, since mean conditional independence fails, synthetic controls cannot effectively identify the treatment effect either (also identifies* 2 *rather than zero).*

Example 11 (treated lie outside convex hull of controls) *Suppose the DGP is the same as example 8 for individuals 1 and 2 but the weights are* $W^T = \begin{bmatrix} 1 & 1 \end{bmatrix}$ *(not a convex combination).*

Y_0	Y	I_1	I_2	I_3	T	$d = I_3 \times T$	$\lambda\mu$
-2	-2	1	0	0	0	0	-3
-2	-2	1	0	0	0	0	-3
4	4	1	0	0	1	0	2
4	4	0	1	0	0	0	3
4	4	0	1	0	0	0	3
0	0	0	1	0	1	0	-2
2	2	0	0	1	0	0	0
2	2	0	0	1	0	0	0
4	6	0	0	1	1	1	0

$\alpha = 6 - 4 = 2$. *D-I-D identifies*

$$E[Y \mid X] = -\frac{1}{3}I_1 + 2\frac{1}{3}I_2 + 2I_3 + T + 3I_3 \times T$$

but the treatment effect is 2 *not* 3. *Since the test subject lies outside the convex hull of the control subjects' predicted outcomes, the synthetic control* *strategy ineffectively identifies the treatment effect.*

	V fixed	*V estimated*
$V\begin{pmatrix} I_1 \\ I_2 \\ \widehat{Y}_i \end{pmatrix}$	$\begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}$	$\begin{bmatrix} 0.44 & 0 & 0 \\ 0 & 0.22 & 0 \\ 0 & 0 & 0.34 \end{bmatrix}$
W^T	$\begin{bmatrix} 0.435 & 0.565 \end{bmatrix} \neq \begin{bmatrix} 1 & 1 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{3} & \frac{2}{3} \end{bmatrix} \neq \begin{bmatrix} 1 & 1 \end{bmatrix}$
$\alpha = Y_{33} - W^T \begin{bmatrix} Y_{13} \\ Y_{23} \end{bmatrix}$	$4.26 \neq 2$	$4\frac{2}{3} \neq 2$

where $\widehat{Y}_i = \sum_{t=1}^{2} k_t Y_{it} = \frac{1}{2}(Y_{i1} + Y_{i2})$. *Hence, the bias in the treatment effect is greater with synthetic controls than the bias for D-I-D.*

3.6 Dynamic treatment effects

Suppose we have a panel data setting in which we observe individuals or firms potentially adopting and leaving (binary) treatment through time. What strategy can be employed to identify average treatment effects in each time period?

3.6.1 Ignorable treatment with unobserved heterogeneity

One approach involves mean conditional independence of the history of treatments given unobserved heterogeneity, c_{i0}, and a set of covariates $\{X_{it} : t = 1, \ldots, T\}$.

$$\begin{aligned} E[Y_{0it} \mid D_i, X_i, c_i] &= E[Y_{0it} \mid X_i, c_i] \\ &= \alpha_{t0} + X_{it}\gamma_0 + c_{i0} \end{aligned}$$

and

$$\begin{aligned} E[Y_{1it} \mid D_i, X_i, c_i] &= E[Y_{1it} \mid X_i, c_i] \\ &= E[Y_{0it} \mid X_i, c_i] + \eta_t + X_{it}\gamma_1 \end{aligned}$$

where $D_i = \{D_{i1}, D_{i2}, \ldots D_{iT}\}$, the entire sequence of treatments.[8]

Average treatment effects are identified via fixed effects. That is, estimate via fixed effects

$$E[Y_{it} \mid D_i, X_i, c_i] = \Im(t)\,\alpha_{t0} + X_{it}\gamma_0 + \Im(i)\,c_{i0} + \Im(t)\,D_{it}\tau_t + D_{it}\left(X_{it} - \overline{X}_t\right)\gamma_1$$

[8] This discussion follows Wooldridge [2010].

where $\Im(i)$ is an individual or firm fixed effect indicator function that captures heterogeneity, c_{i0}, $\Im(t)$ is a time fixed effect indicator function that captures variation in treatment effects through time, $\tau_t = \eta_t + \overline{X}_t\gamma_1$, and $\overline{X}_t$ is the cross-sectional sample average of X during time t.[9]

In other words, panel data accommodate time-varying or dynamic treatment effects. The cost of dynamic strategies is greater pressure on common support as common support issues arise for each time period. Otherwise, we settle for static interpretation of treatment effects where common support is evaluated over the entire time frame.

Next, we explore some examples involving homogeneous treatment effects, heterogeneous treatment effects (in two variations), and limited common support.

3.6.2 Examples

Example 12 (homogeneous treatment effects) *Suppose the DGP is*

t	i	α_t	D_{it}	X_{it}	c_{i0}	η_t	Y_{0it}	Y_{1it}	Y_{it}	TE_{it}
1	1	1	0	−3	−2	1	−4	−9	−4	−5
1	2	1	0	0	2	1	3	4	3	1
1	3	1	0	2	0	1	3	8	3	5
1	4	1	1	−3	−2	1	−4	−9	−9	−5
1	5	1	1	0	2	1	3	4	4	1
1	6	1	1	2	0	1	3	8	8	5
2	1	2	0	−1	−2	3	−1	0	−1	1
2	2	2	0	0	2	3	4	7	4	3
2	3	2	0	3	0	3	5	14	5	9
2	4	2	1	−1	−2	3	−1	0	0	1
2	5	2	1	0	2	3	4	7	7	3
2	6	2	1	3	0	3	5	14	14	9
3	1	1	0	−3	−2	1	−4	−9	−4	−5
3	2	1	0	0	2	1	3	4	3	1
3	3	1	0	5	0	1	6	17	3	11
3	4	1	1	−3	−2	1	−4	−9	−9	−5
3	5	1	1	0	2	1	3	4	4	1
3	6	1	1	5	0	1	6	17	17	11

where $\gamma_0 = 1$, $\gamma_1 = 2$, *and observed outcome is* $Y_{it} = D_{it}Y_{1it} + (1 - D_{it})\,Y_{0it}$. *Average treatment effects vary through time but are homogeneous across*

[9] The average treatment effects are also identified via the fixed effects strategy if $D_{it}X_{it}$ is employed in place of $D_{it}\left(X_{it} - \overline{X}_t\right)$.

treatment subpopulations.

t	$ATE\,(t)$	$ATT\,(t)$	$ATUT\,(t)$
1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
2	$4\frac{1}{3}$	$4\frac{1}{3}$	$4\frac{1}{3}$
3	$2\frac{1}{3}$	$2\frac{1}{3}$	$2\frac{1}{3}$
unconditional	$2\frac{1}{3}$	$2\frac{1}{3}$	$2\frac{1}{3}$

Since the individual effects for $i = 1$ *and* 4*,* $i = 2$ *and* 5*, and* $i = 3$ *and* 6 *are the same, we employ a pooled regression with* $\Im\,(t_1)$*,* $\Im\,(t_2)$*,* $\Im\,(t_3)$*,* $D_{it}\Im\,(t_1)$*,* $D_{it}\Im\,(t_2)$*,* $D_{it}\Im\,(t_3)$*,* $\Im\,(i_{14})$*,* $\Im\,(i_{25})$*, plus covariates,* X_{it} *and* $D_{it}\left(X_{it} - \overline{X}_t\right)$*, with* $\Im\,(i_{36})$ *redundant and therefore excluded from the design matrix. A representative sample produces the following parameters.*

variable	$\Im\,(t_1)$	$\Im\,(t_2)$	$\Im\,(t_3)$	$D_{it}\Im\,(t_1)$	$D_{it}\Im\,(t_2)$
parameter	α_1	α_2	α_3	$\tau_1 = \eta_1 + \overline{X}_1\gamma_1$	$\tau_2 = \eta_2 + \overline{X}_2\gamma_1$
value	1	2	1	$\frac{1}{3}$	$4\frac{1}{3}$

variable	$D_{it}\Im\,(t_3)$	$\Im\,(i_{14})$	$\Im\,(i_{25})$	X_{it}	$D_{it}\left(X_{it} - \overline{X}_t\right)$
parameter	$\tau_3 = \eta_3 + \overline{X}_3\gamma_1$	$c_{1,40}$	$c_{2,50}$	γ_0	γ_1
value	$2\frac{1}{3}$	-2	2	1	2

Estimated conditional average treatment effects are

		$estATT\,(t, X)$	$estATUT\,(t, X)$	$estATE\,(t, X)$
t	X_{it}	$= \tau_t +$	$= \tau_t +$	$= \tau_t +$
		$D_{it}\left(X_{it} - \overline{X}_t\right)\gamma_1$	$(1 - D_{it})\left(X_{it} - \overline{X}_t\right)\gamma_1$	$\left(X_{it} - \overline{X}_t\right)\gamma_1$
1	-3	-5	-5	-5
1	0	1	1	1
1	2	5	5	5
2	-1	1	1	1
2	0	3	3	3
2	3	9	9	9
3	-3	-5	-5	-5
3	0	1	1	1
3	5	11	11	11

Iterated expectations gives the estimated time-dependent but covariate-unconditional average treatment effects

$$\begin{aligned}
estATT\left(t\right) &= E_X\left[estATT\left(t,X\right)\right] \\
&= \tau_t + E\left[X_{it} - \overline{X}_t \mid t, D_{it} = 1\right]\gamma_1 \\
&= \tau_t + \frac{\sum\limits_i D_{it}\Im\left(t\right)\left(X_{it} - \overline{X}_t\right)}{\sum\limits_i D_{it}\Im\left(t\right)}\gamma_1 \\
estATT\left(t = 1\right) &= \frac{1}{3} \\
estATT\left(t = 2\right) &= 4\frac{1}{3} \\
estATT\left(t = 3\right) &= 2\frac{1}{3}
\end{aligned}$$

$$\begin{aligned}
estATUT\left(t\right) &= E_X\left[estATUT\left(t,X\right)\right] \\
&= \tau_t + E\left[X_{it} - \overline{X}_t \mid t, D_{it} = 0\right]\gamma_1 \\
&= \tau_t + \frac{\sum\limits_i \left(1 - D_{it}\right)\Im\left(t\right)\left(X_{it} - \overline{X}_t\right)}{\sum\limits_i \left(1 - D_{it}\right)\Im\left(t\right)}\gamma_1 \\
estATUT\left(t = 1\right) &= \frac{1}{3} \\
estATUT\left(t = 2\right) &= 4\frac{1}{3} \\
estATUT\left(t = 3\right) &= 2\frac{1}{3}
\end{aligned}$$

and

$$\begin{aligned}
estATE\left(t\right) &= E_X\left[estATE\left(t,X\right)\right] \\
&= \Pr\left(D = 1 \mid t\right)ATT\left(t\right) + \Pr\left(D = 0 \mid t\right)ATT\left(t\right) \\
&= \tau_t \\
estATE\left(t = 1\right) &= \frac{1}{3} \\
estATE\left(t = 2\right) &= 4\frac{1}{3} \\
estATE\left(t = 3\right) &= 2\frac{1}{3}
\end{aligned}$$

Estimated unconditional or static average treatment effects are

$$\begin{aligned}
estATT &= E_t\left[estATT\left(t\right)\right] \\
&= \sum_t \Pr\left(t\right) estATT\left(t\right) = 2\frac{1}{3} \\
estATUT &= E_t\left[estATT\left(t\right)\right] \\
&= \sum_t \Pr\left(t\right) estATUT\left(t\right) = 2\frac{1}{3} \\
estATE &= E_t\left[estATE\left(t\right)\right] \\
&= \sum_t \Pr\left(t\right) estATE\left(t\right) = 2\frac{1}{3}
\end{aligned}$$

Example 13 (heterogeneous treatment effects) *Suppose everything remains as in example 12 except treatment adoption, D_{it}. The DGP is*

t	i	α_t	D_{it}	X_{it}	c_{i0}	η_t	Y_{0it}	Y_{1it}	Y_{it}	TE_{it}
1	1	1	1	−3	−2	1	−4	−9	−9	−5
1	2	1	0	0	2	1	3	4	3	1
1	3	1	0	2	0	1	3	8	3	5
1	4	1	0	−3	−2	1	−4	−9	−4	−5
1	5	1	0	0	2	1	3	4	3	1
1	6	1	1	2	0	1	3	8	8	5
2	1	2	0	−1	−2	3	−1	0	−1	1
2	2	2	0	0	2	3	4	7	4	3
2	3	2	0	3	0	3	5	14	5	9
2	4	2	1	−1	−2	3	−1	0	0	1
2	5	2	1	0	2	3	4	7	7	3
2	6	2	1	3	0	3	5	14	14	9
3	1	1	0	−3	−2	1	−4	−9	−4	−5
3	2	1	1	0	2	1	3	4	4	1
3	3	1	1	5	0	1	6	17	17	11
3	4	1	1	−3	−2	1	−4	−9	−9	−5
3	5	1	1	0	2	1	3	4	4	1
3	6	1	0	5	0	1	6	17	6	11

Period 2 involves full common support. However, periods 1 and 3 lack common support for $X_{it} = 0$ so even a fully representative sample leaves treatment effects conditional on $X_{it} = 0$ unidentified. Of course, this also impacts any average treatment effects in which we might be interested. Therefore, all inferences are limited to local support for $t = 1$ or 3. Estimation based on a representative sample of all the data produces the same parameter estimates as in example 12 and the same parameter estimates as drawn from a representative sample of the common support data (that is, excluding draws involving $X_{i1} = X_{i3} = 0$). Estimated average treatment effects *based on a representative common support sample (full sample) are*

t	$estATE\,(t, X_{it})$ $(estATE\,(t))$	$estATT\,(t, X_{it})$ $(estATT\,(t))$	$estATUT\,(t, X_{it})$ $(estATUT\,(t))$	*common* *support*
1	0	0	0	$X_{i1} = -3, 2$
1	$\left(\frac{1}{3}\right)$	(0)	$\left(\frac{1}{2}\right)$	
2	$4\frac{1}{3}$	$4\frac{1}{3}$	$4\frac{1}{3}$	*full*
3	3	3	3	$X_{i3} = -3, 5$
3	$\left(2\frac{1}{3}\right)$	(2)	(3)	

	$estATE\,(X_{it})$ $(estATE)$	$estATT\,(X_{it})$ $(estATT)$	$estATUT\,(X_{it})$ $(estATUT)$
local average	$2\frac{5}{7}$	$2\frac{5}{7}$	$2\frac{5}{7}$
unconditional average	$\left(2\frac{1}{3}\right)$	$\left(2\frac{1}{3}\right)$	$\left(2\frac{1}{3}\right)$

Hence, treatment effects are not only time-dependent but also heterogeneous with respect to treatment subpopulation. However, dynamic heterogeneity is not identified because of lack of common support. Suppressing the time dependence of treatment effects, the unconditional or static average treatment effects enjoy full common support as depicted in parentheses of the last row of the above table.

Example 14 (heterogeneity revisited) *Suppose we enrich the above DGP with two more individuals observed at $X_{it} = 0$ one treated and one untreated in each period then the data enjoy full dynamic common support in each* *period.*

t	i	α_t	D_{it}	X_{it}	c_{i0}	η_t	Y_{0it}	Y_{1it}	Y_{it}	TE_{it}
1	1	1	1	-3	-2	1	-4	-9	-9	-5
1	2	1	0	0	2	1	3	4	3	1
1	3	1	0	2	0	1	3	8	3	5
1	4	1	0	-3	-2	1	-4	-9	-4	-5
1	5	1	0	0	2	1	3	4	3	1
1	6	1	1	2	0	1	3	8	8	5
1	7	1	1	0	2	1	3	4	4	1
1	8	1	0	0	2	1	3	4	3	1
2	1	2	0	-1	-2	3	-1	0	-1	1
2	2	2	0	0	2	3	4	7	4	3
2	3	2	0	3	0	3	5	14	5	9
2	4	2	1	-1	-2	3	-1	0	0	1
2	5	2	1	0	2	3	4	7	7	3
2	6	2	1	3	0	3	5	14	14	9
2	7	2	1	0	2	1	4	7	7	3
2	8	2	0	0	2	1	4	7	4	3
3	1	1	0	-3	-2	1	-4	-9	-4	-5
3	2	1	1	0	2	1	3	4	4	1
3	3	1	1	5	0	1	6	17	17	11
3	4	1	1	-3	-2	1	-4	-9	-9	-5
3	5	1	1	0	2	1	3	4	4	1
3	6	1	0	5	0	1	6	17	6	11
3	7	1	1	0	2	1	3	4	4	1
3	8	1	0	0	2	1	3	4	3	1

Average treatment effects vary through time and time-dependent treatment effects are heterogeneous across treatment subpopulations.

t	$ATE\,(t)$	$ATT\,(t)$	$ATUT\,(t)$
1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{3}{5}$
2	4	4	4
3	2	$1\frac{4}{5}$	$2\frac{1}{3}$
unconditional	$2\frac{1}{6}$	$2\frac{1}{6}$	$2\frac{1}{6}$

We again pool parameters as individuals 2, 5, 7, and 8 involve the same heterogeneity, $c_{i0} = 2$. A representative sample produces the following pa-

rameters.

variable	$\Im(t_1)$	$\Im(t_2)$	$\Im(t_3)$	$D_{it}\Im(t_1)$	$D_{it}\Im(t_2)$	$D_{it}\Im(t_3)$
parameter	α_1	α_2	α_3	τ_1	τ_2	τ_3
value	1	2	1	$\frac{1}{2}$	4	2

variable	$\Im(i_{14})$	$\Im(i_{2578})$	X_{it}	$D_{it}\left(X_{it} - \overline{X}_t\right)$
parameter	$c_{1,40}$	$c_{2,50}$	γ_0	γ_1
value	-2	2	1	2

Estimated conditional (on time and covariate) average treatment effects are the same as example 12. Iterated expectations gives the estimated time-dependent but covariate-unconditional average treatment effects. These time-dependent average effects are heterogeneous across treatment subpopulations.

$$\begin{aligned}
estATT(t) &= E_X\left[estATT(t, X)\right] \\
&= \tau_t + E\left[X_{it} - \overline{X}_t \mid t, D_{it} = 1\right]\gamma_1 \\
&= \tau_t + \frac{\sum_i D_{it}\Im(t)\left(X_{it} - \overline{X}_t\right)}{\sum_i D_{it}\Im(t)}\gamma_1 \\
estATT(t = 1) &= \frac{1}{3} \\
estATT(t = 2) &= 4 \\
estATT(t = 3) &= 1\frac{4}{5}
\end{aligned}$$

$$\begin{aligned}
estATUT(t) &= E_X\left[estATUT(t, X)\right] \\
&= \tau_t + E\left[X_{it} - \overline{X}_t \mid t, D_{it} = 0\right]\gamma_1 \\
&= \tau_t + \frac{\sum_i (1 - D_{it})\Im(t)\left(X_{it} - \overline{X}_t\right)}{\sum_i (1 - D_{it})\Im(t)}\gamma_1 \\
estATUT(t = 1) &= \frac{3}{5} \\
estATUT(t = 2) &= 4 \\
estATUT(t = 3) &= 2\frac{1}{3}
\end{aligned}$$

and

$$\begin{aligned}
estATE\left(t\right) &= E_X\left[estATE\left(t,X\right)\right] \\
&= \Pr\left(D=1\mid t\right)ATT\left(t\right)+\Pr\left(D=0\mid t\right)ATT\left(t\right) \\
&= \tau_t \\
estATE\left(t=1\right) &= \frac{1}{2} \\
estATE\left(t=2\right) &= 4 \\
estATE\left(t=3\right) &= 2
\end{aligned}$$

Estimated unconditional or static average treatment effects are

$$\begin{aligned}
estATT &= E_t\left[estATT\left(t\right)\right] \\
&= \sum_t \Pr\left(t\right)estATT\left(t\right)=2\frac{1}{6} \\
estATUT &= E_t\left[estATT\left(t\right)\right] \\
&= \sum_t \Pr\left(t\right)estATUT\left(t\right)=2\frac{1}{6} \\
estATE &= E_t\left[estATE\left(t\right)\right] \\
&= \sum_t \Pr\left(t\right)estATE\left(t\right)=2\frac{1}{6}
\end{aligned}$$

Example 15 (limited common support) *Consider the following modified DGP.*

t	i	α_t	D_{it}	X_{it}	c_{i0}	η_t	Y_{0it}	Y_{1it}	Y_{it}	TE_{it}
1	1	1	1	−3	−2	1	−4	−9	−4	−5
1	2	1	0	0	2	1	3	4	3	1
1	3	1	0	8	0	1	9	26	3	17
1	4	1	0	−3	−2	1	−4	−9	−9	−5
1	5	1	0	0	2	1	3	4	4	1
1	6	1	0	2	0	1	3	8	8	5
2	1	2	0	−1	−2	3	−1	0	−1	1
2	2	2	0	0	2	3	4	7	4	3
2	3	2	0	3	0	3	5	14	5	9
2	4	2	1	−1	−2	3	−1	0	0	1
2	5	2	1	0	2	3	4	7	7	3
2	6	2	1	3	0	3	5	14	14	9
3	1	1	0	−3	−2	1	−4	−9	−4	−5
3	2	1	1	0	2	1	3	4	3	1
3	3	1	1	5	0	1	6	17	3	11
3	4	1	1	−3	−2	1	−4	−9	−9	−5
3	5	1	1	0	2	1	3	4	4	1
3	6	1	1	5	0	1	6	17	17	11

Ignoring common support, a representative sample produces the following parameter estimates

variable	$\Im(t_1)$	$\Im(t_2)$	$\Im(t_3)$	$D_{it}\Im(t_1)$	$D_{it}\Im(t_2)$	$D_{it}\Im(t_3)$
parameter	α_1	α_2	α_3	τ_1	τ_2	τ_3
value	1	2	1	$2\frac{1}{3}$	$4\frac{1}{3}$	$2\frac{1}{3}$

variable	$\Im(i_{14})$	$\Im(i_{25})$	X_{it}	$D_{it}\left(X_{it}-\overline{X}_t\right)$
parameter	$c_{1,40}$	$c_{2,5,7,80}$	γ_0	γ_1
value	-2	2	1	2

and unidentified (due to lack of support) average treatment effects

$$\begin{aligned}
estATT(t) &= E_X\left[estATT(t,X)\right] \\
&= \tau_t + E\left[X_{it}-\overline{X}_t \mid t, D_{it}=1\right]\gamma_1 \\
&= \tau_t + \frac{\sum\limits_i D_{it}\Im(t)\left(X_{it}-\overline{X}_t\right)}{\sum\limits_i D_{it}\Im(t)}\gamma_1 \\
estATT(t=1) &= -5 \\
estATT(t=2) &= 4\frac{1}{3} \\
estATT(t=3) &= 3\frac{4}{5}
\end{aligned}$$

$$\begin{aligned}
estATUT(t) &= E_X\left[estATUT(t,X)\right] \\
&= \tau_t + E\left[X_{it}-\overline{X}_t \mid t, D_{it}=0\right]\gamma_1 \\
&= \tau_t + \frac{\sum\limits_i (1-D_{it})\Im(t)\left(X_{it}-\overline{X}_t\right)}{\sum\limits_i (1-D_{it})\Im(t)}\gamma_1 \\
estATUT(t=1) &= 3\frac{4}{5} \\
estATUT(t=2) &= 4\frac{1}{3} \\
estATUT(t=3) &= -5
\end{aligned}$$

and

$$\begin{aligned}
estATE(t) &= E_X\left[estATE(t,X)\right] \\
&= \Pr(D=1 \mid t)\,ATT(t) + \Pr(D=0 \mid t)\,ATT(t) \\
&= \tau_t \\
estATE(t=1) &= 2\frac{1}{3} \\
estATE(t=2) &= 4\frac{1}{3} \\
estATE(t=3) &= 2\frac{1}{3}
\end{aligned}$$

Unidentified (due to lack of common support) estimated unconditional or static average treatment effects are

$$\begin{aligned}
estATT &= E_t\left[estATT\left(t\right)\right] \\
&= \sum_t \Pr\left(t\right) estATT\left(t\right) = 3 \\
estATUT &= E_t\left[estATT\left(t\right)\right] \\
&= \sum_t \Pr\left(t\right) estATUT\left(t\right) = 3 \\
estATE &= E_t\left[estATE\left(t\right)\right] \\
&= \sum_t \Pr\left(t\right) estATE\left(t\right) = 3
\end{aligned}$$

Common support is challenging in this example. $X_{21} = 0$*,* $X_{51} = 0$*,* $X_{23} = 0$*, and* $X_{53} = 0$ *lack dynamic common support but satisfy static common support, while* $X_{31} = 8$*,* $X_{61} = 2$*,* $X_{33} = 5$*, and* $X_{63} = 5$ *lack both dynamic and static common support. Fixed effects estimation based on a representative sample satisfying dynamic common support results in redundancy of* $\Im\left(i_{14}\right)$*, consequently it is dropped. The parameter estimates are*

variable	$\Im\left(t_1\right)$	$\Im\left(t_2\right)$	$\Im\left(t_3\right)$	$D_{it}\Im\left(t_1\right)$	$D_{it}\Im\left(t_2\right)$	$D_{it}\Im\left(t_3\right)$
parameter	α_1	α_2	α_3	τ_1	τ_2	τ_3
value	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	-5	$4\frac{1}{3}$	-5

variable	$\Im\left(i_{25}\right)$	X_{it}	$D_{it}\left(X_{it} - \overline{X}_t\right)$
parameter	$c_{7,80}$	γ_0	γ_1
value	$3\frac{1}{2}$	$1\frac{1}{2}$	2

and average treatment effects

$$\begin{aligned}
estATT\left(t\right) &= E_X\left[estATT\left(t, X\right)\right] \\
&= \tau_t + E\left[X_{it} - \overline{X}_t \mid t, D_{it} = 1\right]\gamma_1 \\
&= \tau_t + \frac{\sum_i D_{it}\Im\left(t\right)\left(X_{it} - \overline{X}_t\right)}{\sum_i D_{it}\Im\left(t\right)}\gamma_1 \\
estATT\left(t = 1, X_{i1} = -3\right) &= -5 \\
estATT\left(t = 2\right) &= 4\frac{1}{3} \\
estATT\left(t = 3, X_{i3} = -3\right) &= -5
\end{aligned}$$

$$\begin{aligned}
estATUT\left(t\right) &= E_X\left[estATUT\left(t,X\right)\right] \\
&= \tau_t + E\left[X_{it}-\overline{X}_t \mid t, D_{it}=0\right]\gamma_1 \\
&= \tau_t + \frac{\sum\limits_i \left(1-D_{it}\right)\Im\left(t\right)\left(X_{it}-\overline{X}_t\right)}{\sum\limits_i \left(1-D_{it}\right)\Im\left(t\right)}\gamma_1 \\
estATUT\left(t=1,X_{i1}=-3\right) &= -5 \\
estATUT\left(t=2\right) &= 4\frac{1}{3} \\
estATUT\left(t=3,X_{i3}=-3\right) &= -5
\end{aligned}$$

and

$$\begin{aligned}
estATE\left(t\right) &= E_X\left[estATE\left(t,X\right)\right] \\
&= \Pr\left(D=1\mid t\right)ATT\left(t\right)+\Pr\left(D=0\mid t\right)ATT\left(t\right) \\
&= \tau_t \\
estATE\left(t=1,X_{i1}=-3\right) &= -5 \\
estATE\left(t=2\right) &= 4\frac{1}{3} \\
estATE\left(t=3,X_{i3}=-3\right) &= -5
\end{aligned}$$

Estimated unconditional or static average treatment effects based on the dynamic common support data are

$$\begin{aligned}
& estATT\left(X_{i1}=-3,X_{i2}=-1,X_{i2}=0,X_{i2}=3,X_{i3}=-3\right) \\
= \;& E_t\left[estATT\left(t\right)\right] \\
= \;& \sum_t \Pr\left(t\right)estATT\left(t\right)=\frac{3}{5} \\
& estATUT\left(X_{i1}=-3,X_{i2}=-1,X_{i2}=0,X_{i2}=3,X_{i3}=-3\right) \\
= \;& E_t\left[estATT\left(t\right)\right] \\
= \;& \sum_t \Pr\left(t\right)estATUT\left(t\right)=\frac{3}{5} \\
& estATE\left(X_{i1}=-3,X_{i2}=-1,X_{i2}=0,X_{i2}=3,X_{i3}=-3\right) \\
= \;& E_t\left[estATE\left(t\right)\right] \\
= \;& \sum_t \Pr\left(t\right)estATE\left(t\right)=\frac{3}{5}
\end{aligned}$$

Alternatively, estimated unconditional or static average treatment effects based on the static common support data (that is, employing observations *involving $X_{it} = 0$) are*

$$\begin{array}{rcc}
estATT\left(\begin{array}{c} X_{i1}=-3, X_{i1}=0, X_{i2}=-1, \\ X_{i2}=0, X_{i2}=3, X_{i3}=-3, X_{i3}=0 \end{array}\right) & = & \frac{5}{7} \\
estATUT\left(\begin{array}{c} X_{i1}=-3, X_{i1}=0, X_{i2}=-1, \\ X_{i2}=0, X_{i2}=3, X_{i3}=-3, X_{i3}=0 \end{array}\right) & = & \frac{5}{7} \\
estATE\left(\begin{array}{c} X_{i1}=-3, X_{i1}=0, X_{i2}=-1, \\ X_{i2}=0, X_{i2}=3, X_{i3}=-3, X_{i3}=0 \end{array}\right) & = & \frac{5}{7}
\end{array}$$

where parameter estimates are

variable	$\Im(t_1)$	$\Im(t_2)$	$\Im(t_3)$	$D_{it}\Im(t_1)$	$D_{it}\Im(t_2)$	$D_{it}\Im(t_3)$
parameter	α_1	α_2	α_3	τ_1	τ_2	τ_3
value	1	2	1	-2	$4\frac{1}{3}$	-2

variable	$\Im(i_{14})$	$\Im(i_{25})$	X_{it}	$D_{it}\left(X_{it}-\overline{X}_t\right)$
parameter	$c_{1,40}$	$c_{2,7,80}$	γ_0	γ_1
value	-2	2	1	2

If one were to erroneously (due to lack of dynamic common support) attempt to infer dynamic treatment effects from a representative sample the results are heterogeneous.

$$\begin{aligned}
estATT(t) &= E_X\left[estATT(t,X)\right] \\
&= \tau_t + E\left[X_{it}-\overline{X}_t \mid t, D_{it}=1\right]\gamma_1 \\
&= \tau_t + \frac{\sum_i D_{it}\Im(t)\left(X_{it}-\overline{X}_t\right)}{\sum_i D_{it}\Im(t)}\gamma_1 \\
estATT(t=1) &= -5 \\
estATT(t=2) &= 4\frac{1}{3} \\
estATT(t=3) &= -1
\end{aligned}$$

$$\begin{aligned}
estATUT(t) &= E_X\left[estATUT(t,X)\right] \\
&= \tau_t + E\left[X_{it}-\overline{X}_t \mid t, D_{it}=0\right]\gamma_1 \\
&= \tau_t + \frac{\sum_i (1-D_{it})\Im(t)\left(X_{it}-\overline{X}_t\right)}{\sum_i (1-D_{it})\Im(t)}\gamma_1 \\
estATUT(t=1) &= -1 \\
estATUT(t=2) &= 4\frac{1}{3} \\
estATUT(t=3) &= -5
\end{aligned}$$

and

$$\begin{aligned}
estATE\left(t\right) &= E_X\left[estATE\left(t,X\right)\right] \\
&= \Pr\left(D=1\mid t\right)ATT\left(t\right)+\Pr\left(D=0\mid t\right)ATT\left(t\right) \\
&= \tau_t \\
estATE\left(t=1\right) &= -2 \\
estATE\left(t=2\right) &= 4\frac{1}{3} \\
estATE\left(t=3\right) &= -2
\end{aligned}$$

To summarize, panel data afford time-varying or dynamic treatment effects. The identification strategy discussed above is similar to static ignorable treatment strategies except mean conditional independence involves the entire history of treatment. Inferences regarding dynamic treatment effects impose greater strain on common support as overlapping support applies to each time period. Otherwise, we're left with a static interpretation of treatment effects where the common support demands apply to the entire time frame rather than period-by-period.

3.6.3 Dynamic ignorability identification strategy

Dynamic ignorability is an alternative dynamic treatment effect identification strategy to the ignorable treatment fixed effects with unobservable heterogeneity strategy outlined above. Identification involves a condition plausible in many settings.

$$\Pr\left(D_{it}\mid Y_{0it},Y_{1it},X_i^t\right)=\Pr\left(D_{it}\mid X_i^t\right)$$

where X_i^t is the history of covariates, observed outcomes, and treatments up to time t. In some settings, it's plausible that the past history is so informative that current potential outcomes are not conditionally informative of treatment adoption. Matching on $p\left(X_i^t\right)$ or general matching on X_i^t, propensity score weighting, or (nonparametric) regression can be utilized to identify $ATE\left(X_i^t\right)$, $ATT\left(X_i^t\right)$, and $ATUT\left(X_i^t\right)$.

Consider nonparametric identification with dynamic ignorability (akin to the key identifying condition employed by Rosenbaum and Rubin's propensity score matching). The difference in treated and untreated observable outcomes is

$$E\left[Y_{1it}\mid X_i^t,D_{it}=1\right]-E\left[Y_{0it}\mid X_i^t,D_{it}=0\right]$$

By Bayes rule (iterated expectations) this can be rewritten

$$\begin{aligned}
&\frac{E\left[D_{it}Y_{1it}\mid X_i^t\right]}{\Pr\left(D_{it}=1\mid X_i^t\right)}-\frac{E\left[\left(1-D_{it}\right)Y_{0it}\mid X_i^t\right]}{\Pr\left(D_{it}=0\mid X_i^t\right)} \\
= &\frac{E\left[D_{it}\mid Y_{1it},X_i^t\right]E\left[Y_{1it}\mid X_i^t\right]}{\Pr\left(D_{it}=1\mid X_i^t\right)}-\frac{E\left[\left(1-D_{it}\right)\mid Y_{0it},X_i^t\right]E\left[Y_{0it}\mid X_i^t\right]}{\Pr\left(D_{it}=0\mid X_i^t\right)}
\end{aligned}$$

Dynamic ignorability leads to

$$\begin{aligned} & \frac{E\left[D_{it} \mid X_i^t\right] E\left[Y_{1it} \mid X_i^t\right]}{\Pr\left(D_{it}=1 \mid X_i^t\right)} - \frac{E\left[(1-D_{it}) \mid X_i^t\right] E\left[Y_{0it} \mid X_i^t\right]}{\Pr\left(D_{it}=0 \mid X_i^t\right)} \\ = \; & E\left[Y_{1it} - Y_{0it} \mid X_i^t\right] = ATE\left(X_i^t\right) \end{aligned}$$

By an analogous argument average treatment on the treated and untreated are also identified. The counterfactuals are

$$\begin{aligned} E\left[Y_{0it} \mid X_i^t, D_{it}=1\right] &= \frac{E\left[D_{it} \mid Y_{0it}, X_i^t\right] E\left[Y_{0it} \mid X_i^t\right]}{\Pr\left(D_{it}=1 \mid X_i^t\right)} \\ &= \frac{E\left[D_{it} \mid X_i^t\right] E\left[Y_{0it} \mid X_i^t\right]}{\Pr\left(D_{it}=1 \mid X_i^t\right)} \\ &= E\left[Y_{0it} \mid X_i^t\right] \end{aligned}$$

and

$$\begin{aligned} E\left[Y_{1it} \mid X_i^t, D_{it}=0\right] &= \frac{E\left[(1-D_{it}) \mid Y_{1it}, X_i^t\right] E\left[Y_{1it} \mid X_i^t\right]}{\Pr\left(D_{it}=0 \mid X_i^t\right)} \\ &= \frac{E\left[(1-D_{it}) \mid X_i^t\right] E\left[Y_{1it} \mid X_i^t\right]}{\Pr\left(D_{it}=0 \mid X_i^t\right)} \\ &= E\left[Y_{1it} \mid X_i^t\right] \end{aligned}$$

The above identification strategy applies to nonparametric regression (regression adjustment) and matching approaches. For treatment on treated and untreated, unconditional expectations are determined by creating matches (for matching strategies) on the treated and untreated subsamples, respectively, and summing over X_i^t or $p\left(X_i^t\right) \equiv \Pr\left(D_{it}=1 \mid X_i^t\right)$ for propensity score matching.

Propensity score weighting strategies follow a similar development.

$$\begin{aligned} & E\left[\left(D_{it} - p\left(X_i^t\right)\right) Y_{it} \mid X_i^t\right] \\ = \; & E\left[\begin{array}{c} \left(D_{it} - p\left(X_i^t\right)\right) \\ \times \left(D_{it} Y_{1it} + (1-D_{it}) Y_{0it}\right) \mid X_i^t \end{array}\right] \end{aligned}$$

Expanding gives

$$E\left[\left(D_{it} Y_{1it} - p\left(X_i^t\right) D_{it} Y_{1it} - p\left(X_i^t\right)(1-D_{it}) Y_{0it}\right) \mid X_i^t\right]$$

where $D_{it}(1-D_{it}) = 0$. Factoring constants, $p\left(X_i^t\right)$, outside the expectation operation and collecting terms leads to

$$\begin{aligned} & E\left[D_{it} Y_{1it} \mid X_i^t\right] - p\left(X_i^t\right) E\left[D_{it} Y_{1it} \mid X_i^t\right] - p\left(X_i^t\right) E\left[(1-D_{it}) Y_{0it} \mid X_i^t\right] \\ = \; & \left(1 - p\left(X_i^t\right)\right) E\left[D_{it} Y_{1it} \mid X_i^t\right] - p\left(X_i^t\right)\left(E\left[Y_{0it} \mid X_i^t\right] - E\left[D_{it} Y_{0it} \mid X_i^t\right]\right) \end{aligned}$$

Applying iterated expectations gives

$$\begin{array}{rl} = & \left(1-p\left(X_i^t\right)\right)\left\{ \begin{array}{c} \Pr\left(D_{it}=1\mid X_i^t\right)E\left[D_{it}Y_{1it}\mid X_i^t,D_{it}=1\right] \\ +\Pr\left(D_{it}=0\mid X_i^t\right)E\left[D_{it}Y_{1it}\mid X_i^t,D_{it}=0\right] \end{array} \right\} \\ & -p\left(X_i^t\right)\left(\begin{array}{c} E\left[Y_{0it}\mid X_i^t\right]-\Pr\left(D_{it}=1\mid X_i^t\right)E\left[D_{it}Y_{0it}\mid X_i^t,D_{it}=1\right] \\ -\Pr\left(D_{it}=0\mid X_i^t\right)E\left[D_{it}Y_{0it}\mid X_i^t,D_{it}=0\right] \end{array} \right) \end{array}$$

Recognizing $\Pr\left(D_{it}=1\mid X_i^t\right)\equiv p\left(X_i^t\right)$ and $E\left[D_{it}Y_{kit}\mid X_i^t,D_{it}=0\right]=0$ for $k=0$ or 1 leads to

$$p\left(X_i^t\right)\left(1-p\left(X_i^t\right)\right)E\left[Y_{1it}\mid X_i^t,D_{it}=1\right]-p\left(X_i^t\right)\left(1-p\left(X_i^t\right)\right)E\left[Y_{0it}\mid X_i^t,D_{it}=1\right]$$

Employing mean conditional independence of potential outcomes with and without treatment yields

$$\begin{array}{rl} & p\left(X_i^t\right)\left(1-p\left(X_i^t\right)\right)E\left[Y_{1it}\mid X_i^t\right]-p\left(X_i^t\right)\left(1-p\left(X_i^t\right)\right)E\left[Y_{0it}\mid X_i^t\right] \\ = & p\left(X_i^t\right)\left(1-p\left(X_i^t\right)\right)E\left[Y_{1it}-Y_{0it}\mid X_i^t\right] \\ = & p\left(X_i^t\right)\left(1-p\left(X_i^t\right)\right)ATE\left(X_i^t\right) \end{array}$$

Hence,

$$ATE\left(X_i^t\right)=\frac{E\left[\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}\mid X_i^t\right]}{p\left(X_i^t\right)\left(1-p\left(X_i^t\right)\right)}$$

An alternative identification demonstration with multiple combined steps is below.

$$\begin{array}{rl} & E\left[\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}\mid X_i^t\right] \\ = & E\left[\begin{array}{c} \left(D_{it}-p\left(X_i^t\right)\right) \\ \times\left(D_{it}Y_{1it}+\left(1-D_{it}\right)Y_{0it}\right) \end{array} \mid X_i^t \right] \\ = & \Pr\left(D_{it}=1\mid X_i^t\right)\times \\ & E\left[\begin{array}{c} \left(D_{it}-p\left(X_i^t\right)\right) \\ \times\left(D_{it}Y_{1it}+\left(1-D_{it}\right)Y_{0it}\right) \end{array} \mid X_i^t,D_{it}=1 \right] \\ & +\Pr\left(D_{it}=0\mid X_i^t\right)\times \\ & E\left[\begin{array}{c} \left(D_{it}-p\left(X_i^t\right)\right) \\ \times\left(D_{it}Y_{1it}+\left(1-D_{it}\right)Y_{0it}\right) \end{array} \mid X_i^t,D_{it}=0 \right] \\ = & \Pr\left(D_{it}=1\mid X_i^t\right)\times \\ & E\left[\begin{array}{c} \left(1-p\left(X_i^t\right)\right) \\ \times\left(1Y_{1it}+\left(1-1\right)Y_{0it}\right) \end{array} \mid X_i^t,D_{it}=1 \right] \\ & +\Pr\left(D_{it}=0\mid X_i^t\right)\times \\ & E\left[\begin{array}{c} \left(0-p\left(X_i^t\right)\right) \\ \times\left(0Y_{1it}+\left(1-0\right)Y_{0it}\right) \end{array} \mid X_i^t,D_{it}=0 \right] \end{array}$$

Hence,

$$\begin{aligned}
& E\left[\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}\mid X_i^t\right] \\
&= \Pr\left(D_{it}=1\mid X_i^t\right)\left(1-p\left(X_i^t\right)\right)E\left[Y_{1it}\mid X_i^t,D_{it}=1\right] \\
&\quad -p\left(X_i^t\right)\Pr\left(D_{it}=0\mid X_i^t\right)E\left[Y_{0it}\mid X_i^t,D_{it}=0\right] \\
&= p\left(X_i^t\right)\left(1-p\left(X_i^t\right)\right)E\left[Y_{1it}-Y_{0it}\mid X_i^t\right] \\
&= p\left(X_i^t\right)\left(1-p\left(X_i^t\right)\right)ATE\left(X_i^t\right) \\
ATE\left(X_i^t\right) &= \frac{E\left[\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}\mid X_i^t\right]}{p\left(X_i^t\right)\left(1-p\left(X_i^t\right)\right)}
\end{aligned}$$

Dynamic ignorability produces the following treatment-subpopulation estimands.

$$\begin{aligned}
E\left[\frac{\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}}{1-p\left(X_i^t\right)}\mid X_i^t\right] &= p\left(X_i^t\right)E\left[Y_{1it}-Y_{0it}\mid X_i^t\right] \\
&= \Pr\left(D_{it}=1\mid X_i^t,Y_{1it}\right)E\left[Y_{1it}\mid X_i^t\right] \\
&\quad -\Pr\left(D_{it}=1\mid X_i^t,Y_{0it}\right)E\left[Y_{0it}\mid X_i^t\right] \\
&= E\left[D_{it}\left(Y_{1it}-Y_{0it}\right)\mid X_i^t\right] \\
&= p\left(X_i^t\right)E\left[Y_{1it}-Y_{0it}\mid X_i^t,D_{it}=1\right] \\
&= p\left(X_i^t\right)ATT\left(X_i^t\right) \\
ATT\left(X_i^t\right) &= E\left[\frac{\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}}{1-p\left(X_i^t\right)}\mid X_i^t\right]/p\left(X_i^t\right)
\end{aligned}$$

and

$$\begin{aligned}
E\left[\frac{\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}}{p\left(X_i^t\right)}\mid X_i^t\right] &= \left(1-p\left(X_i^t\right)\right)E\left[Y_{1it}-Y_{0it}\mid X_i^t\right] \\
&= \Pr\left(D_{it}=0\mid X_i^t,Y_{1it}\right)E\left[Y_{1it}\mid X_i^t\right] \\
&\quad -\Pr\left(D_{it}=0\mid X_i^t,Y_{0it}\right)E\left[Y_{0it}\mid X_i^t\right] \\
&= E\left[\left(1-D_{it}\right)\left(Y_{1it}-Y_{0it}\right)\mid X_i^t\right] \\
&= \left(1-p\left(X_i^t\right)\right)E\left[Y_{1it}-Y_{0it}\mid X_i^t,D_{it}=0\right] \\
&= \left(1-p\left(X_i^t\right)\right)ATUT\left(X_i^t\right) \\
ATUT\left(X_i^t\right) &= E\left[\frac{\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}}{p\left(X_i^t\right)}\mid X_i^t\right]/\left(1-p\left(X_i^t\right)\right)
\end{aligned}$$

Iterated expectations leads to unconditional average treatment effects.

$$\begin{aligned}
ATE &= E_X\left[E\left[\frac{\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}}{p\left(X_i^t\right)\left(1-p\left(X_i^t\right)\right)}\mid X_i^t\right]\right] \\
&= E\left[\frac{\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}}{p\left(X_i^t\right)\left(1-p\left(X_i^t\right)\right)}\right]
\end{aligned}$$

$$\begin{aligned}
E_X\left[E\left[\frac{\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}}{1-p\left(X_i^t\right)}\mid X_i^t\right]\right] &= E_X\left[p\left(X_i^t\right)ATT\left(X_i^t\right)\right] \\
&= E_X\left[D_{it}ATT\mid X_i^t\right] \\
E\left[\frac{\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}}{1-p\left(X_i^t\right)}\right] &= \Pr\left(D_{it}=1\right)ATT \\
ATT &= E\left[\frac{\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}}{1-p\left(X_i^t\right)}\right]/\Pr\left(D_{it}=1\right)
\end{aligned}$$

and

$$\begin{aligned}
E_X\left[E\left[\frac{\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}}{p\left(X_i^t\right)}\mid X_i^t\right]\right] &= E_X\left[\left(1-p\left(X_i^t\right)\right)ATUT\left(X_i^t\right)\right] \\
&= E_X\left[\left(1-D_{it}\right)ATUT\mid X_i^t\right] \\
E\left[\frac{\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}}{p\left(X_i^t\right)}\right] &= \Pr\left(D_{it}=0\right)ATUT \\
ATUT &= E\left[\frac{\left(D_{it}-p\left(X_i^t\right)\right)Y_{it}}{p\left(X_i^t\right)}\right]/\Pr\left(D_{it}=0\right)
\end{aligned}$$

However, in settings where once adopted treatment is rarely dropped, common support is likely to be extremely limited or even nonexistent. This trade-off between satisfying ignorability and overlapping support is common to all treatment effect identification strategies (for instance, more regressors eases ignorability but challenges support and vice versa) but seems especially perplexing with dynamic ignorability.

Next, we explore instrumental variable strategies.

3.7 Local average treatment effects

The above example suggests the conditions for ignorable treatment may severely limit identification and estimation of treatment effects. A common complementary approach to expanding the set of regressors (ignorable treatment) is to employ instrumental variables. *Instrumental variables*, Z, are variables that are associated with treatment choice, D, but unrelated to the outcomes with and without treatment, Y_1 and Y_0. The idea is we can manipulate treatment choice with the instrument but leave outcomes unaffected. This permits extrapolation from observables to counterfactuals, $E\left[Y_1\mid D=0\right]$ and $E\left[Y_0\mid D=1\right]$.

If outcomes with treatment, Y_1, and outcomes without treatment, Y_0, are independent of a binary instrument, Z, then the discrete marginal treatment effect or local average treatment effect,

$$LATE=E\left[Y_1-Y_0\mid D_1-D_0=1\right]$$

where $D_1 = (D \mid Z = 1)$ and $D_0 = (D \mid Z = 0)$ equals

$$\frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[D \mid Z = 1] - E[D \mid Z = 0]}$$

This quantity (ratio) can be estimated from observables, therefore *LATE* is identified. In fact, this quantity (*estimand*) is estimated by standard two-stage instrumental variable estimation (*2SLS-IV*).

3.7.1 2SLS-IV estimation

Suppose we envision the regression in error form

$$Y = \alpha + \beta D + \varepsilon$$

but $E[\varepsilon \mid D] \neq 0$, then *OLS* provides inconsistent parameter estimates but instrumental variable estimation can rectify the problem. As the name suggests, *2SLS-IV* estimation involves two stages of projections. The first stage puts the explanatory variables of interest (here, treatment, D) in the columns of the instruments, Z. In other words, we construct[10]

$$\begin{aligned} \widehat{D} &= Z\left(Z^T Z\right)^{-1} Z^T \widetilde{D} \\ &= P_Z \widetilde{D} \end{aligned}$$

where $\widetilde{D} = D - \overline{D}$ is the estimated mean deviation. Then, we estimate

$$E[Y \mid D] = a + b\widehat{D}$$

where the estimate of β is

$$b = \left(X^T X\right)^{-1} X^T \widetilde{Y}$$

and $X = \widehat{D}$ and $\widetilde{Y} = Y - \overline{Y}$. Since

$$\begin{aligned} \left(X^T X\right)^{-1} X^T \widetilde{Y} &= \left(\widetilde{D}^T P_Z P_Z \widetilde{D}\right)^{-1} \widetilde{D}^T P_Z P_Z \widetilde{Y} \\ &= \left(P_Z \widetilde{D}\right)^{-1} \left(\widetilde{D}^T P_Z\right)^{-1} \widetilde{D}^T P_Z P_Z \widetilde{Y} \\ &= \left(P_Z \widetilde{D}\right)^{-1} P_Z \widetilde{Y} \\ &= \frac{\frac{1}{n} Z^T \widetilde{Y}}{\frac{1}{n} Z^T \widetilde{D}} \end{aligned}$$

$\frac{\frac{1}{n} Z^T \widetilde{Y}}{\frac{1}{n} Z^T \widetilde{D}}$ estimates $\frac{E\left[\widetilde{Y} \mid Z=1\right] - E\left[\widetilde{Y} \mid Z=0\right]}{E\left[\widetilde{D} \mid Z=1\right] - E\left[\widetilde{D} \mid Z=0\right]} = \frac{E[Y \mid Z=1] - E[Y \mid Z=0]}{E[D \mid Z=1] - E[D \mid Z=0]} = LATE$. It's time for an example.

[10] To simplify matters, we work with a single variable by utilizing mean deviations of all variables. We discuss *2SLS-IV* estimation more generally in the appendix to this chapter.

3.7.2 IV example 1

Suppose the *DGP* is

Y	D	Y_1	Y_0	Z
15	1	15	10	1
15	1	15	10	0
10	1	10	10	1
10	0	10	10	0
10	0	5	10	1
10	0	5	10	0

IV example 1: $LATE = 0$

If we estimate by *OLS* we find

$$E[Y \mid D] = 10 + 3\frac{1}{3}D$$

suggesting the average treatment effect is $3\frac{1}{3}$. As treatment is not ignorable, this is a false conclusion,

$$ATE = E[Y_1 - Y_0] = 10 - 10 = 0$$

Now, if we think of the first two rows as state 1 and successive pairs of rows similarly where treatment, D, is potentially manipulated via the instrument, Z, then we can estimate *LATE* via *2SLS-IV*. With this *DGP*, *LATE* is identified for only state 2 (rows 3 and 4) since $D_1 - D_0 = 1$ (the compliers — individuals induced to select treatment when $Z = 1$ but not when $Z = 0$). State 1 represents individuals who always select treatment and state 3 represents individuals who never select treatment. Clearly, $LATE = \frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]} = \frac{10-10}{2/3-1/3} = 0$ and *2SLS-IV* estimates $\frac{\frac{1}{n}Z^T\widetilde{Y}}{\frac{1}{n}Z^T\widetilde{D}} = 0$, in large samples. Hence, for this *DGP* and instrument, $LATE = ATE$. but we should not expect this, in general, as the next examples illustrate.

3.7.3 IV example 2

Suppose the *DGP* is

Y	D	Y_1	Y_0	Z
15	1	15	10	1
10	0	15	10	0
10	1	10	10	1
10	1	10	10	0
10	0	5	10	1
10	0	5	10	0

IV example 2: $LATE = 5$

OLS estimates

$$E[Y \mid D] = 10 + 1\frac{2}{3}D$$

which again fails to identify the average treatment effect, $ATE = 0$. Now, the compliers are reflected by state 1 alone, and $LATE = 5$ while ATE continues to be zero. Also, *2SLS-IV* estimates $\frac{\frac{1}{n}Z^T\widetilde{Y}}{\frac{1}{n}Z^T\widetilde{D}} = 5$ in large samples.

3.7.4 IV example 3

The DGP along with the instrument identifies the particular marginal treatment effect. Consider another variation

Y	D	Y_1	Y_0	Z
15	1	15	10	1
15	1	15	10	0
10	0	10	10	1
10	0	10	10	0
5	1	5	10	1
10	0	5	10	0

IV example 3: $LATE = -5$

OLS again supplies an inconsistent estimate of *ATE*.

$$E[Y \mid D] = 10 + 1\frac{2}{3}D$$

As the compliers are individuals in state 3, $LATE = -5$ and *2SLS-IV* estimates $\frac{\frac{1}{n}Z^T\widetilde{Y}}{\frac{1}{n}Z^T\widetilde{D}} = -5$ in large samples.

3.7.5 IV example 4

Sometimes $LATE$ equals the average treatment effect on the treated. If no one adopts treatment when the instrument value equals zero, then $LATE = ATT$. Consider the DGP

Y	D	Y_1	Y_0	Z
15	1	15	10	1
10	0	15	10	0
20	0	20	20	1
20	0	20	20	0
10	1	10	10	1
10	0	10	10	0

IV example 4: $LATE = ATT$

OLS estimates

$$E[Y \mid D] = 15 - 2.5D$$

but $ATE = 1\frac{2}{3}$ (opposite directions, or a Simpson's paradox result) and $ATUT = 1.25$. $LATE = 2.5$ is defined by states 1 and 3 and since $\Pr(D = 1 \mid Z = 0) = 0$, $LATE = ATT = E[Y_1 - Y_0 \mid D = 1] = 2.5$. And, *2SLS-IV* estimates $\frac{\frac{1}{n}Z^T\widetilde{Y}}{\frac{1}{n}Z^T\widetilde{D}} = 2.5$ in large samples.

3.7.6 IV example 5

LATE equals the average treatment effect on the untreated if everyone adopts treatment when the instrument equals unity. Consider the *DGP*

Y	D	Y_1	Y_0	Z
15	1	15	10	1
10	0	15	10	0
20	1	20	10	1
20	1	20	10	0
10	1	10	10	1
10	0	10	10	0

IV example 5: $LATE = ATUT$

OLS estimates

$$E[Y \mid D] = 10 + 6.25D$$

but $ATE = 5$ and $ATT = 6.25$. $LATE = 2.5$ is defined by states 1 and 3 and since $\Pr(D = 1 \mid Z = 1) = 1$, $LATE = ATUT = E[Y_1 - Y_0 \mid D = 1] = 2.5$. And, *2SLS-IV* estimates $\frac{\frac{1}{n}Z^T\widetilde{Y}}{\frac{1}{n}Z^T\widetilde{D}} = 2.5$ in large samples.

3.7.7 IV example 6

Unfortunately, if some individuals are induced to accept treatment when the instrument changes to one but others are induced to move away from treatment with the same instrumental variable manipulation, then extant instrumental variable strategies break down. Uniformity is a condition for *IV* identification of *LATE*, any defiers result in treatment effect identification failure. We illustrate the problem once again with a simple binary instrument. Consider the *DGP*

Y	D	Y_1	Y_0	Z
10	1	10	5	1
5	0	10	5	0
10	1	10	5	1
10	1	10	5	0
10	1	10	5	1
5	0	10	5	0
−5	0	10	−5	1
10	1	10	−5	0

IV example 6: defiers

OLS estimates

$$E[Y \mid D] = 1\frac{2}{3} + 8\frac{1}{3}D$$

but $ATE = 7.5$, $ATT = 7$, and $ATUT = 8\frac{1}{3}$. $LATE = 5$ is defined by states 1 and 3 but state 4 violates uniformity. *2SLS-IV* estimates $\frac{\frac{1}{n}Z^T\widetilde{Y}}{\frac{1}{n}Z^T\widetilde{D}} = -5$ in large samples, a gross misstatement of the treatment effect as the treatment effect for every individual is positive!

This illustrates the trouble two-way flows cause in the identification of treatment effects. Extant *IV* strategies rely on uniformity either toward treatment or away from treatment by all individuals, not some individuals toward and others away from treatment in response to changes in the instrument. Further, this simple binary instrumental variable strategy identifies the local average treatment effect for an unidentified subpopulation of compliers. Nonetheless, binary *IV* identifies marginal treatment effects for this subpopulation, a parameter surely of some interest.

3.7.8 Linear IV with covariates

What treatment effect does linear two-stage least squares instrumental variables (*2SLS-IV*) identify when covariates are present? We'll continue with a binary instrument, Z, where *2SLS-IV* identifies *LATE*, an average marginal treatment effect, in the absence of covariates. However, in the presence of covariates *2SLS-IV* identifies *LATE* only for a special case. Let potential outcomes be

$$Y_j = X\beta_j + V_j, \quad j = 0, 1$$

We present two examples to illustrate special cases. In the first case, treatment effects are homogeneous and *2SLS-IV* as well as *OLS* identify the treatment effect. In the second case, treatment effects are heterogeneous but $E[V_j Z \mid X_k = 1] = 0$ and $E[V_j \mid X_k = 1] = 0$ for $j = 0, 1$ and all x_k where X denotes combinations of regressors such that $X_k = 1$ when combination k of regressor values is the row in play and zero otherwise (in other words, the regressors are treated as K mutually exclusive indicator variables and X is a fixed design matrix). These conditions allow $LATE(X_k = 1)$ to be identified by *2SLS-IV* conditional on $X_k = 1$ but not by *OLS*.

A third example illustrates a more general case in which *2SLS-IV* does not identify $LATE(X_k = 1)$ or *LATE*. *2SLS-IV* identifies a weighted average of effects defined by *IV* estimation conditional on $X_k = 1$ where

$$E_X[E[V_j Z \mid X_k = 1]] = E[V_j Z] = 0$$

and

$$E_X[E[V_j \mid X_k = 1]] = E[V_j] = 0$$

for $j = 0, 1$ but

$$E[V_j Z \mid X_k = 1] \neq 0$$

for some x_k and $j = 0$ or 1. We've written these more general conditions in iterated expectation form to emphasize example two is a special case.

The treatment effect identified via *2SLS-IV* is

$$\begin{aligned}\gamma &= \frac{E_X\left[E\left[Y \cdot \left(E\left[D \mid X, Z\right] - E\left[D \mid X\right]\right)\right] \mid X_k = 1\right]}{E_X\left[E\left[D \cdot \left(E\left[D \mid X, Z\right] - E\left[D \mid X\right]\right)\right] \mid X_k = 1\right]} \\ &= \frac{E_X\left[\omega\left(X_k = 1\right)\gamma\left(X_k = 1\right)\right]}{E_X\left[\omega\left(X_k = 1\right)\right]} \\ &= \frac{E\left[Y \cdot \left(E\left[D \mid X, Z\right] - E\left[D \mid X\right]\right)\right]}{E\left[D \cdot \left(E\left[D \mid X, Z\right] - E\left[D \mid X\right]\right)\right]}\end{aligned}$$

where

$$\omega\left(X_k = 1\right) = E\left[E\left[D \mid X, Z\right] \cdot \left(E\left[D \mid X, Z\right] - E\left[D \mid X\right]\right) \mid X_k = 1\right]$$

are the weights and

$$\gamma\left(X_k = 1\right) = \frac{E\left[Y \cdot \left(E\left[D \mid X, Z\right] - E\left[D \mid X\right]\right) \mid X_k = 1\right]}{E\left[D \cdot \left(E\left[D \mid X, Z\right] - E\left[D \mid X\right]\right) \mid X_k = 1\right]}$$

are the *2SLS-IV* effects identified at each $X_k = 1$.[11] Hence, the *2SLS-IV* treatment effect depends on the instrumental variables and the covariates.

Example 16 (homogeneous treatment effect) *Suppose we have the following DGP*

Y	Y_1	Y_0	X_1	X_2	X_3	D	V_1	V_0	Z
6	6	4	1	0	0	1	1	3	1
4	6	4	1	0	0	0	1	3	0
4	4	-2	1	0	0	1	-1	-3	1
-2	4	-2	1	0	0	0	-1	-3	0
4	4	4	0	1	0	1	-2	2	1
4	4	4	0	1	0	0	-2	2	0
8	8	0	0	1	0	1	2	-2	1
0	8	0	0	1	0	0	2	-2	0
10	10	4	0	0	1	1	3	1	1
4	10	4	0	0	1	0	3	1	0
4	4	2	0	0	1	1	-3	-1	1
2	4	2	0	0	1	0	-3	-1	0

where $Y = DY_1 + (1 - D)Y_0$ and importantly, ignorable treatment is satisfied as $E[Y_j \mid X, D] = E[Y_j \mid X]$ for all X_k and $j = 0, 1$ so that OLS identifies the homogeneous average treatment effect as well as $E[V_j Z \mid X_k = 1] =$

[11] See Angrist and Imbens [1995] for an even more general treatment.

0 and $E[V_j \mid X_k = 1] = 0$ *for* $j = 0, 1$ *and all* X_k *is satisfied so that 2SLS-IV identifies the same homogeneous treatment effect. Various conditional and unconditional average treatment effects are as follows.*

	conditional			
TE	$X_1 = 1$	$X_2 = 1$	$X_3 = 1$	*unconditional*
OLS	4	4	4	4
$LATE$	4	4	4	4
$2SLS - IV$	4	4	4	4
$\omega(X_k = 1)$	0.25	0.25	0.25	
ATT	4	4	4	4
$ATUT$	4	4	4	4
ATE	4	4	4	4

Average treatment effects are all equal including effects identified by $2SLS-IV$, $LATE$, *and* OLS.

Example 17 (*2SLS-IV* effect = *LATE*) *Suppose we have the following DGP*

Y	Y_1	Y_0	X_1	X_2	X_3	D	V_1	V_0	Z
6	6	4	1	0	0	1	1	3	1
4	6	4	1	0	0	0	1	3	0
4	4	−2	1	0	0	1	−1	−3	1
4	4	−2	1	0	0	1	−1	−3	0
8	8	4	0	1	0	1	2	2	1
4	8	4	0	1	0	0	2	2	0
0	4	0	0	1	0	0	−2	−2	1
0	4	0	0	1	0	0	−2	−2	0
10	10	4	0	0	1	1	3	1	1
4	10	4	0	0	1	0	3	1	0
2	4	2	0	0	1	0	−3	−1	1
2	4	2	0	0	1	0	−3	−1	0

where $E[V_j Z \mid X_k = 1] = 0$ *and* $E[V_j \mid X_k = 1] = 0$ *for* $j = 0, 1$ *and all* X_k *is satisfied so that 2SLS-IV identifies the same average treatment effect as LATE(*$X_k = 1$*) and LATE. Various conditional and unconditional average treatment effects are as follows.*

	conditional			
TE	$X_1 = 1$	$X_2 = 1$	$X_3 = 1$	*unconditional*
OLS	0.6667	6.6667	7.3333	4.1143
$LATE$	2	4	6	4
$2SLS - IV$	2	4	6	4
$\omega(X_k = 1)$	0.0625	0.0625	0.0625	
ATT	4.6667	4	6	4.8
$ATUT$	2	4	3.3333	3.4286
ATE	4	4	4	4

Average treatment effects identified by $2SLS-IV$ *equal* $LATE$ *but* OLS *fails to identify any (standard) average treatment effect.*

Example 18 (*2SLS-IV* effect $\neq$ *LATE*) *Suppose we have the following DGP*

Y	Y_1	Y_0	X_1	X_2	X_3	D	V_1	V_0	Z
6	6	4	1	0	0	1	1	3	1
4	6	4	1	0	0	0	1	3	0
8	8	0	1	0	0	1	3	-1	0
8	8	0	1	0	0	1	3	-1	0
8	8	4	0	1	0	1	2	2	1
4	8	4	0	1	0	0	2	2	0
-1	6	-1	0	1	0	0	0	-3	1
-1	6	-1	0	1	0	0	0	-3	1
4	4	4	0	0	1	1	-3	1	1
4	4	4	0	0	1	0	-3	1	0
1	4	1	0	0	1	0	-3	-2	0
1	4	1	0	0	1	0	-3	-2	0

where $E[V_jZ \mid X_k = 1] \neq 0$ *and* $E[V_j \mid X_k = 1] \neq 0$ *but* $E[V_jZ] = 0$ *and* $E[V_j] = 0$ *for* $j = 0, 1$ *is satisfied so that 2SLS-IV identifies a different average treatment effect from* $LATE(X_k = 1)$ *and LATE. Various conditional and unconditional average treatment effects are as follows.*

	conditional			
TE	$X_1 = 1$	$X_2 = 1$	$X_3 = 1$	*unconditional*
OLS	3.3333	7.3333	2	5.0857
$LATE$	2	4	0	2
$2SLS-IV$	-2	-6	2	0.9091
$\omega(X_k = 1)$	0.02083	0.02083	0.1875	
ATT	6	4	0	4.4
$ATUT$	2	6	2	3.7143
ATE	5	5.5	1.5	4

Average treatment effects identified by $2SLS-IV$ *are unequal to* $LATE$ *or any other (standard) average treatment effect.*

3.8 Treatment effects and control functions

Another approach that may be effective for identifying treatment effects utilizes control functions. That is, functions which directly control the source of selection bias. Consider a simple data generating process (to keep the discussion compact, there are no regressors).

$$Y_j = \mu_j + V_j, \quad j = 0, 1$$

where μ_j is the mean of outcome and V_j is the unobserved (not residual) portion of outcome for treatment j.

Y	D	Y_1	Y_0	V_1	V_0
15	1	15	9	3	-1
14	1	14	10	2	-2
13	1	13	11	1	-3
13	0	11	13	-1	3
14	0	10	14	-2	2
15	0	9	15	-3	1

If we attempt to estimate average treatment effects via an exogenous dummy variable regression[12]

$$E[Y \mid D] = \mu_0 + (\mu_1 - \mu_0) D$$

we find that OLS estimates

$$E[Y \mid D] = 14 + 0D$$

Suggesting all average treatment effects are zero. While it is the case, the unconditional average treatment effect is zero

$$E[Y_1 - Y_0] = 12 - 12 = 0$$

the means for outcome with treatment and with no treatment are not identified as OLS suggests the mean of each is 14 while the DGP clearly indicates the mean of each is 12. Further, we may have more interest in the average treatment effect on the treated and untreated but OLS does not identify either of these quantities. The fundamental problem is that the basic condition for a well-posed regression, $E[V_j \mid X] = 0$, is not satisfied. Rather,

$$E[V_1 \mid D = 1] = \frac{1}{3}(3 + 2 + 1) = 2$$

$$E[V_1 \mid D = 0] = \frac{1}{3}(-3 - 2 - 1) = -2$$

$$E[V_0 \mid D = 1] = \frac{1}{3}(-3 - 2 - 1) = -2$$

and

$$E[V_0 \mid D = 0] = \frac{1}{3}(3 + 2 + 1) = 2$$

Nonetheless, the average treatment effects on the treated (ATT) and untreated ($ATUT$) are well-defined.

$$\begin{aligned} ATT &= E[Y_1 \mid D = 1] - E[Y_0 \mid D = 1] \\ &= 12 + 2 - (12 - 2) \\ &= 4 \end{aligned}$$

[12] This is in the same spirit as a single factor $ANOVA$ with binary factor levels.

and

$$\begin{aligned} ATUT &= E[Y_1 \mid D=0] - E[Y_0 \mid D=0] \\ &= (12-2) - (12+2) \\ &= -4 \end{aligned}$$

Also, these quantities readily connect to the average treatment effect.

$$\begin{aligned} ATE &= \Pr(D=1)\, ATT + (1-\Pr(D=1))\, ATUT \\ &= \frac{1}{2}(4) + \frac{1}{2}(-4) = 0 \end{aligned}$$

The key is to determine a path from observable data to these quantities. The control function approach attempts to include functions in the regression that control for the source of selection bias, $E[V_j \mid D]$. Then, the means can be properly identified and estimation from observable data is feasible.

The most popular control function approach was developed by Nobel laureate, James Heckman. Briefly, the idea is treatment selection by an individual reflects expected utility maximizing behavior. The data analyst (manager, social scientist, etc.) observes some factors influencing this choice but other factors are unobserved (by the analyst). These unobserved components lead to a stochastic process description of individual choice behavior. The key to this stochastic description is the probability assignment to the unobservable component. Heckman argues when the probability assignment is Gaussian or normal, then we can treat the problem as a truncated regression exercise. And, when common support conditions for the regressors are satisfied, in principle, average treatment effects on the treated, untreated, and the unconditional average are identified. Otherwise, when common support conditions are limited, local average treatment effects only are identified. We sketch the ideas below and relate them to the above example.[13]

3.8.1 Inverse Mills control function strategy

Consider the DGP where choice is represented by a latent variable characterizing the difference in expected utility associated with treatment or no treatment, observed choice, and outcome equations with treatment and without treatment.

latent choice equation:

$$D^* = W\theta + V_D$$

observed choice:

$$D = \begin{matrix} 1 & \text{if } D^* > 0 \quad V_D > -W\theta \\ 0 & \text{otherwise} \end{matrix}$$

[13] This subsection is heavily laden with notation — bear with us.

outcome equations:

$$\begin{aligned} Y_1 &= \mu_1 + X\beta_1 + V_1 \\ Y_0 &= \mu_0 + X\beta_0 + V_0 \end{aligned}$$

Heckman's two-stage estimation procedure is as follows. First, estimate θ via a probit regression of D on $W = \{\iota, X, Z\}$ where Z is an instrumental variable and identify observations with common support (that is, observations for which the regressors, X, for the treated overlap with regressors for the untreated). Second, regress Y onto

$$\left\{\iota, D, X, D\left(X - E\left[X\right]\right), D\left(-\frac{\phi}{\Phi}\right), \left(1 - D\right)\frac{\phi}{1 - \Phi}\right\}$$

for the overlapping subsample. With full support, the coefficient on D is a consistent estimator of ATE; with less than full common support, we have a local average treatment effect.[14]

Wooldridge suggests identification of

$$ATE = \mu_1 - \mu_0 + E\left[X\right]\left(\beta_1 - \beta_0\right)$$

via α in the regression

$$\begin{aligned} E\left[Y \mid X, Z\right] &= \mu_0 + \alpha D + X\beta_0 + D\left(X - E\left[X\right]\right)\left(\beta_1 - \beta_0\right) \\ &\quad -D\rho_{1V_D}\sigma_1\frac{\phi\left(W\theta\right)}{\Phi\left(W\theta\right)} + \left(1 - D\right)\rho_{0V_D}\sigma_0\frac{\phi\left(W\theta\right)}{1 - \Phi\left(W\theta\right)} \end{aligned}$$

This follows from the observable response

$$\begin{aligned} Y &= D\left(Y_1 \mid D = 1\right) + \left(1 - D\right)\left(Y_0 \mid D = 0\right) \\ &= \left(Y_0 \mid D = 0\right) + D\left[\left(Y_1 \mid D = 1\right) - \left(Y_0 \mid D = 0\right)\right] \end{aligned}$$

and applying conditional expectations

$$\begin{aligned} E\left[Y_1 \mid X, D = 1\right] &= \mu_1 + X\beta_1 - \rho_{1V_D}\sigma_1\frac{\phi\left(W\theta\right)}{\Phi\left(W\theta\right)} \\ E\left[Y_0 \mid X, D = 0\right] &= \mu_0 + X\beta_0 + \rho_{0V_D}\sigma_0\frac{\phi\left(W\theta\right)}{1 - \Phi\left(W\theta\right)} \end{aligned}$$

[14] We should point out here that second stage OLS does not provide valid estimates of standard errors. As Heckman points out there are two additional concerns: the errors are heteroskedastic (so an adjustment such as White suggested is needed) and θ has to be estimated (so we must account for this added variation). Heckman identifies a valid variance estimator for this two-stage procedure.

Simplification produces Wooldridge's result.

$$
\begin{aligned}
E[Y \mid X, Z] &= E[(Y_0 \mid D=0) + D\{(Y_1 \mid D=1) - (Y_0 \mid D=0)\} \mid X, Z] \\
&= \mu_0 + X\beta_0 + \rho_{0V_D}\sigma_0 \frac{\phi(W\theta)}{1-\Phi(W\theta)} \\
&\quad +D\left(\mu_1 + X\beta_1 - \rho_{1V_D}\sigma_1 \frac{\phi(W\theta)}{\Phi(W\theta)}\right) \\
&\quad -D\left(\mu_0 + X\beta_0 + \rho_{0V_D}\sigma_0 \frac{\phi(W\theta)}{1-\Phi(W\theta)}\right)
\end{aligned}
$$

now rearrange terms

$$
\begin{aligned}
&\mu_0 + D\{\mu_1 - \mu_0 + E[X](\beta_1 - \beta_0)\} + X\beta_0 + D(X - E[X])(\beta_1 - \beta_0) \\
&-D\rho_{1V_D}\sigma_1 \frac{\phi(W\theta)}{\Phi(W\theta)} + (1-D)\rho_{0V_D}\sigma_0 \frac{\phi(W\theta)}{1-\Phi(W\theta)}
\end{aligned}
$$

The coefficient on D, $\{\mu_1 - \mu_0 + E[X](\beta_1 - \beta_0)\}$, is ATE.

The key ideas behind treatment effect identification via control functions can be illustrated by reference to this case.

$$E[Y_j \mid X, D=j] = \mu_j + X\beta_j + E[V_j \mid D=j]$$

Given the conditions, $E[V_j \mid D=j] \neq 0$ unless $Corr(V_j, V_D) = \rho_{jV_D} = 0$. For $\rho_{jV_D} \neq 0$,

$$E[V_1 \mid D=1] = \rho_{1V_D}\sigma_1 E[V_D \mid V_D > -W\theta]$$

$$E[V_0 \mid D=1] = \rho_{0V_D}\sigma_0 E[V_D \mid V_D > -W\theta]$$

$$E[V_1 \mid D=0] = \rho_{1V_D}\sigma_1 E[V_D \mid V_D \leq -W\theta]$$

and

$$E[V_0 \mid D=0] = \rho_{0V_D}\sigma_0 E[V_D \mid V_D \leq -W\theta]$$

The final term in each expression is the expected value of a truncated standard normal random variate where

$$h_1 \equiv E[V_D \mid V_D > -W\theta] = E[V_D \mid V_D < W\theta] = -\frac{\phi(W\theta)}{\Phi(W\theta)}$$

and

$$h_0 \equiv E[V_D \mid V_D \leq -Z\theta] = E[V_D \mid V_D \geq Z\theta] = \frac{\phi(W\theta)}{1-\Phi(W\theta)}$$

Putting this together, we have

$$E[Y_1 \mid X, D=1] = \mu_1 + X\beta_1 - \rho_{1V_D}\sigma_1 \frac{\phi(W\theta)}{\Phi(W\theta)}$$

$$E\left[Y_0 \mid X, D=0\right] = \mu_0 + X\beta_0 + \rho_{0V_D}\sigma_0 \frac{\phi\left(W\theta\right)}{1-\Phi\left(W\theta\right)}$$

and counterfactuals

$$E\left[Y_0 \mid X, D=1\right] = \mu_0 + X\beta_0 - \rho_{0V_D}\sigma_0 \frac{\phi\left(W\theta\right)}{\Phi\left(W\theta\right)}$$

and

$$E\left[Y_1 \mid X, D=0\right] = \mu_1 + X\beta_1 + \rho_{1V_D}\sigma_1 \frac{\phi\left(W\theta\right)}{1-\Phi\left(W\theta\right)}$$

The appeal of Heckman's inverse Mills ratio strategy can be seen in its estimation simplicity and the ease with which treatment effects are then identified. Of course, this doesn't justify the identification conditions — only our understanding of the data can do that. The conditional average treatment effect on the treated is

$$ATT\left(X,Z\right) = \mu_1 - \mu_0 + X\left(\beta_1 - \beta_0\right) - \left(\rho_{1V_D}\sigma_1 - \rho_{0V_D}\sigma_0\right)\frac{\phi\left(W\theta\right)}{\Phi\left(W\theta\right)}$$

and by iterated expectations (with full support), we have the unconditional average treatment effect on the treated

$$ATT = \mu_1 - \mu_0 + E\left[X\right]\left(\beta_1 - \beta_0\right) - \left(\rho_{1V_D}\sigma_1 - \rho_{0V_D}\sigma_0\right) E\left[\frac{\phi\left(W\theta\right)}{\Phi\left(W\theta\right)}\right]$$

Also, the conditional average treatment effect on the untreated is

$$ATUT\left(X,Z\right) = \mu_1 - \mu_0 + X\left(\beta_1 - \beta_0\right) + \left(\rho_{1V_D}\sigma_1 - \rho_{0V_D}\sigma_0\right)\frac{\phi\left(W\theta\right)}{1-\Phi\left(W\theta\right)}$$

and by iterated expectations, we have the unconditional average treatment effect on the untreated

$$ATUT = \mu_1 - \mu_0 + E\left[X\right]\left(\beta_1 - \beta_0\right) + \left(\rho_{1V_D}\sigma_1 - \rho_{0V_D}\sigma_0\right) E\left[\frac{\phi\left(W\theta\right)}{1-\Phi\left(W\theta\right)}\right]$$

Since

$$\begin{aligned} ATE\left(X,Z\right) &= \Pr\left(D=1 \mid X,Z\right) ATT\left(X,Z\right) \\ &\quad + \Pr\left(D=0 \mid X,Z\right) ATUT\left(X,Z\right) \\ &= \Phi\left(W\theta\right) ATT\left(X,Z\right) + \left(1-\Phi\left(W\theta\right)\right) ATUT\left(X,Z\right) \end{aligned}$$

we have the conditional average treatment effect is

$$\begin{aligned} ATE\left(X,Z\right) &= \mu_1 - \mu_0 + X\left(\beta_1 - \beta_0\right) \\ &\quad + \left(\rho_{1V}\sigma_1 - \rho_{0V_D}\sigma_0\right)\phi\left(W\theta\right) - \left(\rho_{1V}\sigma_1 - \rho_{0V_D}\sigma_0\right)\phi\left(W\theta\right) \\ &= \mu_1 - \mu_0 + X\left(\beta_1 - \beta_0\right) \end{aligned}$$

and by iterated expectations, we have the unconditional average treatment effect is

$$ATE = \mu_1 - \mu_0 + E\left[X\right]\left(\beta_1 - \beta_0\right)$$

3.8.2 Back to the example

Now, we return to the example and illustrate this control function strategy. Suppose the first stage probit regression produces the following scaled hazard rates (inverse Mills ratios) where $h = D * h_1 + (1 - D)\, h_0$, $h_1 = -\omega_1 \frac{\phi(W\theta)}{\Phi(W\theta)}$ $h_0 = \omega_0 \frac{\phi(W\theta)}{1-\Phi(W\theta)}$, and $\omega_1 = \omega_0 = 2.843$.[15]

Y	D	Y_1	Y_0	V_1	V_0	h
15	1	15	9	3	-3	-3
14	1	14	10	2	-2	-2
13	1	13	11	1	-1	-1
13	0	11	13	-1	1	1
14	0	10	14	-2	2	2
15	0	9	15	-3	3	3

The large sample second stage regression is

$$E\left[Y \mid D, h\right] = 12 + 0D - 1.0\left(D \times h_1\right) + 1.0\left((1 - D) \times h_0\right)$$

Estimated average treatment effects consistently identify (again, a large sample result) the average treatment effects as follows. The average treatment effect is estimated via the coefficient on D

$$estATE = 0$$

[15] Clearly, we've omitted details associated with the first stage. Suffice to say we have regressors (instruments) related to selection, D, but that are uninformative about outcomes, Y_1 and Y_0 (otherwise we would include them in the output regressions). The instruments, $Z = \left[\begin{array}{cccc} Z_1 & Z_2 & Z_3 & Z_4 \end{array}\right]$ (no intercept; tabulated below) employed are orthogonal to Y_1 and Y_0.

Z_1	Z_2	Z_3	Z_4
5	4	3	1
-6	-5	-4	-2
0	0	0	1
0	0	1	0
0	1	0	0
1	0	0	0

In fact, they form a basis for the nullspace to $\left[\begin{array}{cc} Y_1 & Y_0 \end{array}\right]$. When we return to this setting to explore Bayesian analysis, we'll be more explicit about this first stage estimation but we bypass this stage for now.

Other estimated averages of interest are

$$\begin{aligned}
estE\left[Y_1 \mid D=1\right] &= 12+0-1.0\left(\frac{-3-2-1}{3}\right) \\
&= 14 \\
estE\left[Y_1 \mid D=0\right] &= 12-1.0\left(\frac{3+2+1}{3}\right) \\
&= 10 \\
estE\left[Y_0 \mid D=1\right] &= 12+0+1.0\left(\frac{-3-2-1}{3}\right) \\
&= 10 \\
estE\left[Y_0 \mid D=0\right] &= 12+1.0\left(\frac{3+2+1}{3}\right) \\
&= 14
\end{aligned}$$

Hence, the estimated average treatment effect on the treated is

$$\begin{aligned}
estATT &= estE\left[Y_1 \mid D=1\right]-estE\left[Y_0 \mid D=1\right] \\
&= 14-10=4
\end{aligned}$$

and the estimated average treatment effect on the untreated is

$$\begin{aligned}
estATUT &= estE\left[Y_1 \mid D=0\right]-estE\left[Y_0 \mid D=0\right] \\
&= 10-14=-4
\end{aligned}$$

We see the control function strategy has effectively addressed selection bias and allowed us to identify some average treatment effects of interest even though the DGP poses serious challenges.

3.9 Pursuit of higher explanatory power

A word of caution. Frequently, we utilize explanatory power to help gauge model adequacy. This is a poor strategy in the analysis of treatment effects. Higher explanatory power in either the selection equation or the outcome equations does not ensure identification of average treatment effects. We present two examples below in which higher explanatory power models completely undermine identification of treatment effects.

3.9.1 Outcomes model example

It might be tempting to employ the instrument $Z_5 = \left[\begin{array}{cccccc} 1 & 0 & -1 & -1 & 0 & 1 \end{array}\right]^T$ as a regressor as it perfectly explains observed outcome Y. Estimates are

$$E\left[Y \mid Z_5, D, h\right] = 14 + 1.0Z_5 + 0D + 0\left(D \times h_1\right) + 0\left(\left(1-D\right) \times h_0\right)$$

However, recall our objective is to estimate treatment effects and they draw from outcomes, Y_1 and Y_0, which are only partially observed and Z_5 is independent of these outcomes.[16] This regression produces severe selection bias, disguises endogeneity, suggests homogeneous outcome when it is heterogeneous, and masks self-selection. In other words, it could hardly be more misleading even though it has higher explanatory power.

3.9.2 Selection model example

Suppose we add the regressor,

$$X = \begin{bmatrix} 1 & 0 & 1 & -1 & 0 & -1 \end{bmatrix}^T$$

to the instruments in the selection equation so that the regressors in the probit model are[17]

$$\begin{aligned} W &= \begin{bmatrix} X & Z_1 & Z_2 & Z_3 & Z_4 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 5 & 4 & 3 & 1 \\ 0 & -6 & -5 & -4 & -2 \\ 1 & 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

Again, we suppress probit estimation details. The estimated outcomes model conditional on the "control functions" is

$$E[Y \mid X, D, h] = 14 + 0D + 0(D \times h_1) + 0((1 - D) \times h_0)$$

As in the higher explanatory power outcomes model, this treatment effect identification strategy is a complete bust. Here, it is because the regressor, X, dominates the instruments in explaining treatment choice and it's the instruments that allow manipulation of choice without affecting outcome — the key to identifying properties of the counterfactuals. Hence, the regression is plagued by severe selection bias, disguises endogeneity and heterogeneity of outcomes, and hides self-selection inherent to the setting.

[16] Identification of instruments is extremely delicate because we don't observe a portion of the outcome distributions.

[17] Employment of a perfect predictor, say

$$X_2 = \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}^T$$

is well known to create estimation problems. In this case any positive weight on X_2 supplies an equally good fit and makes any other regressors superfluous in the selection equation. Results for the perfect regressor case parallel that presented, except with the perfect predictor, x_2, the coefficients on the control functions, $(D \times h_1)$ and $((1 - D) \times h_0)$, are actually indeterminant since they are a linear combination of the intercept and D.

To summarize, high explanatory power of either the selection equation or outcome equations does not indicate a well-specified model. As the above examples suggest, higher explanatory power can undermine our treatment effect identification strategy. When addressing counterfactuals and treatment effects, we have no choice but to rely on what we know prior to examining the evidence (namely, theory) in specifying the model.[18]

The foregoing discussion focuses on discrete treatment effects. Prior to segueing to Bayesian analysis, we briefly discuss continuous treatment effects.

3.10 Continuous treatment effects

Suppose we're interested in identifying the average marginal effect of information precision on some response variable, say product market share.[19] The idea is firm managers engage in costly information search to improve firm productivity and competitive advantage. More precise information is beneficial but also more costly. Specifically, benefits are weakly concave while costs are convex. Then, the optimal precision level associated with information discovery is when expected marginal benefits equal expected marginal costs of information precision. Of course, both expected benefits and expected costs depend on the resident circumstance for each firm. Therefore, we must accommodate heterogeneity. If the first order condition (foc) is linear in precision

$$b_i - k_1\tau_i - (mc_i + k_2\tau_i) = 0$$

to yield optimal information precision

$$\tau_i = \frac{b_i - mc_i}{k}$$

where $k = k_1 + k_2$ then the structural model is quadratic in the endogenous variable, on integrating with respect to information precision.[20]

$$Y_i = a_i + b_i\tau_i - \frac{1}{2}k_1\tau^2$$

[18] We don't mean to imply that diagnostic checking based on the evidence is to be shunned. To the contrary, but we must exercise caution and bear in mind how we're exploiting observables to infer unobservables (e.g., counterfactuals).

[19] Interpretation depends on the response variable (see the discussion below).

[20] This setup is very similar to Card's [2001] classic, hedonic partial equilibrium years of schooling model. Numerous accounting variations arise from the simple setting including equilibrium reporting strategies. We'll stay with the current setting to keep it simple and highlight econometric challenges. Notice, like in the discrete choice setting, there is a latent choice model and an outcomes model for benefits of choice that is the focus of our data analysis.

where Y_i is response or outcome, a_i is the firm-specific constant of integration, b_i is firm i's marginal effect of information precision on outcome (when $k_1 = 0$ as below), and τ_i is firm i's information precision choice. To accommodate heterogeneity, both the intercept and slope (or marginal effect of precision) are allowed to vary by firm. This leads to a random coefficients design. As with the years of schooling setting, unobservable heterogeneity (discussed shortly) leads to a correlated random coefficients design.

To simplify discussion of the correlated random coefficients design, we set k_1 equal to zero (this mirrors Heckman and Vytlacil [1998], Wooldridge [2003], and others). With instrumental variables, Z, the model in error form (and with $k_1 = 0$) is

$$Y_i = a_i + b_i \tau_i + e_i, \quad E\left[e_i \mid X_i, Z_i\right] = 0$$

Now, let the intercepts and slopes depend on observables (covariates) X_i

$$a_i = \gamma_0 + X_i \gamma + c_i, \quad E\left[c_i \mid X_i, Z_i\right] = 0$$

and

$$b_i = \beta + \left(X_i - E\left[X\right]\right)\delta + \upsilon_i, \quad E\left[\upsilon_i \mid X_i, Z_i\right] = 0$$

where

$$E\left[Y \mid a, b, \tau, X, Z\right] = E\left[Y \mid a, b, \tau\right],$$

and exclusion restrictions

$$E\left[a \mid X, Z\right] = E\left[a \mid X\right] = \gamma_0 + X\gamma,$$

$$E\left[b \mid X, Z\right] = E\left[b \mid X\right] = \beta + \left(X - E\left[X\right]\right)\delta.$$

Combining yields

$$\begin{aligned} Y_i &= \gamma_0 + X_i\gamma + \left[\beta + \left(X_i - E\left[X\right]\right)\delta\right]\tau_i + c_i + \upsilon_i\tau_i + e_i \\ &= \gamma_0 + X_i\gamma + \tau_i\beta + \left(X_i - E\left[X\right]\right)\delta\tau_i + c_i + \upsilon_i\tau_i + e_i \end{aligned}$$

The challenge resides with interaction of the unobservable and information precision choice, $\upsilon_i\tau_i$. As Heckman and Vytlacil [1998] and Wooldridge [2003] emphasize we don't require $E\left[\upsilon_i\tau_i \mid X_i, Z_i\right] = 0$, rather instrumental variable (IV) identification of average marginal treatment effects is satisfied if $E\left[\upsilon_i\tau_i \mid X_i, Z_i\right] = \alpha$, a constant that doesn't depend on X_i and Z_i.

Interpretation depends on the response variable. For example, if product market share is the response variable then we're looking at benefits excluding costs as in the years of schooling setting.[21] The average treatment effect $E\left[b\right] = \beta$ (ATE) for a random draw from the population

[21] On the other hand, if the response variable is income then we're looking at benefits net of costs. If income is accounting income we have an approximation, a subset of benefits net of a subset of costs, as accounting recognition comes into play.

is the average marginal increase in product market share when information precision is increased. The average treatment effect on the treated $E\left[b_i \mid \tau_i\right] = E\left[b_i \mid b_i = mc_i + k\tau_i\right] = \beta + \left(X_i - E\left[X\right]\right)\delta$ (ATT) indicates the average marginal increase in market share from increasing information precision at $\tau_i = \left(b_i - mc_i\right)/k$, or in other words, at $b_i = mc_i + k\tau_i$. As discussed by Heckman and Vytlacil [1998], this is straightforward when mc_i is stochastically independent of (a_i, b_i).

Next, we explore some simple data generating processes (DGP) to illustrate when OLS and (so-called) forbidden regressions are consistent or otherwise and compare them with two instrumental variables strategies. For purposes of illustration, we consider a firm to be identified by a particular value of report precision, τ.

The **first IV strategy** employs instruments created out of **polynomials** in X and Z. We're interested in $E\left[\tau \mid X, Z\right]$ and $E\left[Y \mid X, Z\right]$ which are closely approximated with linear projections in polynomials of X and Z. In our simple examples we employ instruments $\left\{\iota, X, Z, X^2\right\}$ where ι is a vector of ones (if the rank condition is satisfied we might add XZ as well as higher order terms).[22] Hence, the first-stage regresses τ_i and $\left(X_i - \overline{X}\right)\tau_i$ and the identified instruments and their predicted values, $\widehat{\tau}_i$ and $\widehat{\left(X_i - \overline{X}\right)}\tau_i$, are employed in the second-stage regression of Y_i on $\left\{\iota, X_i, \widehat{\tau}_i, \widehat{\left(X_i - \overline{X}\right)}\tau_i\right\}$.

The **second IV strategy**, employs instruments created from a regression of information precision τ on X and Z, $\widehat{\tau} = L\left[\tau \mid X, Z\right]$ where $L\left[\cdot\right]$ refers to a linear projection of the leading vector on the trailing column space. That is, instruments are $\left\{\iota, X, \widehat{\tau}, \widehat{\tau}X\right\}$ and the first-stage regressions project τ and $\left(X - \overline{X}\right)\tau$ onto these instruments to create predicted values $\widetilde{\tau}$ and $\widetilde{\left(X - \overline{X}\right)}\tau$. Then, the second-stage involves regressing Y_i onto $\left\{\iota, X_i, \widetilde{\tau}_i, \widetilde{\left(X_i - \overline{X}\right)}\tau_i\right\}$.

Contrast these IV strategies with a forbidden regression. A **forbidden regression** employs plug-in regressors, not instruments, $\left\{\iota, X, \widehat{\tau}, \widehat{\tau}\left(X - \overline{X}\right)\right\}$ where $\widehat{\tau}$ is created as in the say the first-stage of the second IV strategy. To emphasize the distinction, recall IV is a two-stage approach. A forbidden regression plugs-in $\widehat{\tau}\left(X - \overline{X}\right)$ in place of $\tau\left(X - \overline{X}\right)$ then regresses outcome on the resultant plug-in regressors. On the other hand, the second IV strategy employs $\widehat{\tau}$ as a plug-in for τ then employs two-stage (not single-stage as utilized by a forbidden regression) IV estimation with the plug-in instruments. Nonlinearity in the endogenous variable may result in inconsistent forbidden regression estimators in settings where IV is consistent. The examples demonstrate this is particularly troublesome for the

[22] In our examples the rank condition is satisifed (and identification results are the same) with either X^2 or XZ but not both as Z is binary.

current setting when $E[XZ] \neq E[XZ \mid Z]$. Of course, there are settings where neither *IV* nor forbidden regression is consistent.

3.10.1 Case 1: all identification strategies are effective

Let $\beta = \gamma = E[b] = 1, \delta = 0.5$, and $\gamma_0 = c = e = 0$. For case 1, $E[\upsilon_i \tau_i \mid X_i, Z_i] = 0$ for all firms, *OLS* identifies the average marginal treatment effect $E[b] = 1$ (*ATE*) as well as the average marginal treatment effect for the treated $E[b_i \mid \tau_i] = \beta + (X_i - E[X])\,\delta$ (*ATT*). Also, since $E[\upsilon_i \tau_i \mid X_i, Z_i]$ is constant and $E[XZ] = E[XZ \mid Z]$, the forbidden regression also identifies both average effects. Not surprisingly, both *IV* strategies identify both average treatment effects. The *DGP* for case 1 is

Y	a	b	X	Z	τ	υ	$\upsilon\tau$
0	-2	2	-2	-1	1	2	2
-4	-2	-2	-2	-1	1	-2	-2
6	1	2.5	1	-1	2	1	2
2	1	0.5	1	-1	2	-1	-2
2.5	-1	$\frac{7}{6}$	-1	1	3	$\frac{2}{3}$	2
-1.5	-1	$-\frac{1}{6}$	-1	1	3	$-\frac{2}{3}$	-2
12	2	2.5	2	1	4	$\frac{1}{2}$	2
8	2	1.5	2	1	4	$-\frac{1}{2}$	-2

Case 1 *DGP*: *OLS, FR & IV* identify treatment effects

The table below reports parameters associated with the *DGP* along with results for various identification strategies: *OLS*, *FR* (forbidden regression), IV_{X^2} (the polynomial *IV* strategy), and $IV_{\widehat{\tau}}$ (the plug-in *IV* strategy).

parameter	*DGP*	*OLS*	*FR*	IV_{X^2}	$IV_{\widehat{\tau}}$
γ_0	0	0	0	0	0
γ	1	1	1	1	1
$\beta = E[b] = ATE$	1	1	1	1	1
δ	0.5	0.5	0.5	0.5	0.5
$E[b \mid \tau = 1] = ATT\,(\tau = 1)$	0	0	0	0	0
$E[b \mid \tau = 2] = ATT\,(\tau = 2)$	1.5	1.5	1.5	1.5	1.5
$E[b \mid \tau = 3] = ATT\,(\tau = 3)$	0.5	0.5	0.5	0.5	0.5
$E[b \mid \tau = 4] = ATT\,(\tau = 4)$	2	2	2	2	2

Case 1 results: *OLS, FR & IV* identify treatment effects

3.10.2 Case 2: forbidden regression (FR) fails

For case 2, $E[\upsilon_i \tau_i \mid X_i, Z_i] = 0$ for all firms but $E[XZ]$ depends on Z ($E[XZ] \neq E[XZ \mid Z]$), hence *OLS* identifies $E[b] = 1$ (*ATE*) as well as $E[b_i \mid \tau_i] = \beta + (X_i - E[X])\,\delta$ (*ATT*) but *FR* fails. Not surprisingly, the *IV* strategies also identify both average treatment effects. The *DGP* for case 2 is

Y	a	b	X	Z	τ	υ	$\upsilon\tau$
23.125	-2	25.125	-2	-1	1	25	25
-26.875	-2	-24.875	-2	-1	1	-25	-25
50.25	-1	25.625	-1	-1	2	25	50
-49.75	-1	-24.375	-1	-1	2	-25	-50
78.375	0	26.125	0	1	3	25	75
-71.625	0	-23.875	0	1	3	-25	-75
110.5	2	27.125	2	1	4	25	100
-89.5	2	-22.875	2	1	4	-25	-100

Case 2 *DGP*: *OLS & IV* identify treatment effects but *FR* fails

The table below reports parameters associated with the *DGP* along with results for various identification strategies: *OLS*, *FR* (forbidden regression), IV_{X^2} (the polynomial *IV* strategy), and $IV_{\widehat{\tau}}$ (the plug-in *IV* strategy).

parameter	*DGP*	*OLS*	*FR*	IV_{X^2}	$IV_{\widehat{\tau}}$
γ_0	0	0	0.4336	0	0
γ	1	1	1.2700	1	1
$\beta = E[b] = ATE$	1	1	1.0143	1	1
δ	0.5	0.5	0.4107	0.5	0.5
$E[b \mid \tau = 1] = ATT(\tau = 1)$	0.125	0.125	0.2955	0.125	0.125
$E[b \mid \tau = 2] = ATT(\tau = 2)$	0.625	0.625	0.7063	0.625	0.625
$E[b \mid \tau = 3] = ATT(\tau = 3)$	1.125	1.125	1.1170	1.125	1.125
$E[b \mid \tau = 4] = ATT(\tau = 4)$	2.125	2.125	1.9384	2.125	2.125

Case 2 results: *OLS & IV* identify treatment effects but *FR* fails

3.10.3 Case 3: OLS fails

For case 3, $E[\upsilon_i \tau_i \mid X_i, Z_i] \neq 0$ but constant for all firms and $E[XZ]$ does not depend on Z, hence *FR* identifies $E[b] = 1$ (*ATE*) as well as $E[b_i \mid \tau_i] = \beta + (X_i - E[X])\,\delta$ (*ATT*) but *OLS* fails. Not surprisingly, the *IV* strategies also identify both average treatment effects. The *DGP* for case 3 is

Y	a	b	X	Z	τ	υ	$\upsilon\tau$
23	-2	25	-2	-1	1	25	25
-52	-2	-25	-2	-1	2	-25	-50
80.5	1	26.5	1	-1	3	25	75
-93	1	-23.5	1	-1	4	-25	-100
126.5	-1	25.5	-1	1	5	25	125
-148	-1	-24.5	-1	1	6	-25	-150
191	2	27	2	1	7	25	175
-182	2	-23	2	1	8	-25	-200

Case 3 *DGP*: *FR & IV* identify treatment effects but *OLS* fails

The table below reports parameters associated with the *DGP* along with results for various identification strategies: *OLS*, *FR* (forbidden regression), IV_{X^2} (the polynomial *IV* strategy), and $IV_{\widehat{\tau}}$ (the plug-in *IV* strategy).

parameter	DGP	OLS	FR	IV_{X^2}	$IV_{\widehat{\tau}}$
γ_0	-12.5	85.6134	-12.5	-12.5	-12.5
γ	1	32.3769	1	1	1
$\beta = E[b] = ATE$	1	-19.4545	1	1	1
δ	0.5	-1.9272	0.5	0.5	0.5
$E[b \mid \tau = 1] = ATT(\tau = 1)$	0	-15.6002	0	0	0
$E[b \mid \tau = 2] = ATT(\tau = 2)$	0	-15.6002	0	0	0
$E[b \mid \tau = 3] = ATT(\tau = 3)$	1.5	-21.3817	1.5	1.5	1.5
$E[b \mid \tau = 4] = ATT(\tau = 4)$	1.5	-21.3817	1.5	1.5	1.5
$E[b \mid \tau = 5] = ATT(\tau = 5)$	0.5	-17.5274	0.5	0.5	0.5
$E[b \mid \tau = 6] = ATT(\tau = 6)$	0.5	-17.5274	0.5	0.5	0.5
$E[b \mid \tau = 7] = ATT(\tau = 7)$	2	-23.3089	2	2	2
$E[b \mid \tau = 8] = ATT(\tau = 8)$	2	-23.3089	2	2	2

Case 3 results: *FR & IV* identify treatment effects but *OLS* fails

3.10.4 Case 4: OLS and FR fail

For case 4, $E[\upsilon_i \tau_i \mid X_i, Z_i] \neq 0$ but constant for all firms and $E[XZ]$ depends on Z, hence *IV* identifies $E[b] = 1$ (ATE) as well as $E[b_i \mid \tau_i] = \beta + (X_i - E[X])\,\delta$ (ATT) but *OLS* and *FR* fail. The *DGP* for case 4 is

Y	a	b	X	Z	τ	υ	$\upsilon\tau$
22.625	-2	24.625	-2	-1	1	25	25
-52	-2	-25.375	-2	-1	2	-25	-50
80.5	-1	25.125	-1	-1	3	25	75
-93	-1	-24.875	-1	-1	4	-25	-100
126.5	2	26.625	2	1	5	25	125
-148	2	-23.375	2	1	6	-25	-150
191	4	27.625	4	1	7	25	175
-182	4	-22.375	4	1	8	-25	-200

Case 4 *DGP*: *IV* identifies treatment effects but *OLS & FR* fail

The table below reports parameters associated with the *DGP* along with results for various identification strategies: *OLS*, *FR* (forbidden regression),

IV_{X^2} (the polynomial IV strategy), and $IV_{\widehat{\tau}}$ (the plug-in IV strategy).

parameter	DGP	OLS	FR	IV_{X^2}	$IV_{\widehat{\tau}}$
γ_0	-12.5	620.8	-9.635	-12.5	-12.5
γ	1	208.1	1.761	1	1
$\beta = E[b] = ATE$	1	-160.1	-0.072	1	1
δ	0.5	-11.582	0.465	0.5	0.5
$E[b \mid \tau = 1] = ATT(\tau = 1)$	-0.375	-15.600	-1.352	-0.375	-0.375
$E[b \mid \tau = 2] = ATT(\tau = 2)$	-0.375	-15.600	-1.352	-0.375	-0.375
$E[b \mid \tau = 3] = ATT(\tau = 3)$	0.125	-21.382	-0.887	0.125	0.125
$E[b \mid \tau = 4] = ATT(\tau = 4)$	0.125	-21.382	-0.887	0.125	0.125
$E[b \mid \tau = 5] = ATT(\tau = 5)$	1.625	-17.527	0.510	1.625	1.625
$E[b \mid \tau = 6] = ATT(\tau = 6)$	1.625	-17.527	0.510	1.625	1.625
$E[b \mid \tau = 7] = ATT(\tau = 7)$	2.625	-23.309	1.440	2.625	2.625
$E[b \mid \tau = 8] = ATT(\tau = 8)$	2.625	-23.309	1.440	2.625	2.625

Case 4 results: *IV* identifies treatment effects but *OLS & FR* fail

3.10.5 Case 5: all identification strategies fail

For case 5, $E[\upsilon_i \tau_i \mid X_i, Z_i] \neq 0$ as well as nonconstant and $E[XZ]$ depends on Z, hence neither IV, OLS or FR identify $E[b] = 1$ (ATE) or $E[b_i \mid \tau_i] = \beta + (X_i - E[X])\,\delta$ (ATT). The DGP for case 5 is

Y	a	b	X	Z	τ	υ	$\upsilon\tau$
38	-2	40	-2	-1	1	40	40
-82	-2	-40	-2	-1	2	-40	-80
120.5	-1	40.5	-1	-1	3	40	120
-159	-1	-39.5	-1	-1	4	-40	-160
130	0	26	0	1	5	25	125
-144	0	-24	0	1	6	-25	-150
195.5	3	27.5	3	1	7	25	175
-177	3	-22.5	3	1	8	-25	-200

Case 5 *DGP*: *IV, OLS & FR* fail

The table below reports parameters associated with the DGP along with results for various identification strategies: OLS, FR (forbidden regression),

IV_{X^2} (the polynomial IV strategy), and $IV_{\hat{\tau}}$ (the plug-in IV strategy).

parameter	DGP	OLS	FR	IV_{X^2}	$IV_{\hat{\tau}}$
γ_0	-16.25	1215	-36.3	-74.375	-74.375
γ	1	524.1	-2.21	-20.563	-20.563
$\beta = E[b] = ATE$	1	-231.3	5.521	12.25	12.25
δ	0.5	-46.05	0.417	2.375	2.375
$E[b \mid \tau = 1] = ATT(\tau = 1)$	0	-139.2	4.688	7.5	7.5
$E[b \mid \tau = 2] = ATT(\tau = 2)$	0	-139.2	4.688	7.5	7.5
$E[b \mid \tau = 3] = ATT(\tau = 3)$	0.5	-185.3	5.104	9.875	9.875
$E[b \mid \tau = 4] = ATT(\tau = 4)$	0.5	-185.3	5.104	9.875	9.875
$E[b \mid \tau = 5] = ATT(\tau = 5)$	1	-231.3	5.521	12.25	12.25
$E[b \mid \tau = 6] = ATT(\tau = 6)$	1	-231.3	5.521	12.25	12.25
$E[b \mid \tau = 7] = ATT(\tau = 7)$	2.5	-369.5	6.771	19.375	19.375
$E[b \mid \tau = 8] = ATT(\tau = 8)$	2.5	-369.5	6.771	19.375	19.375

Case 5 results: *IV*, *OLS & FR* fail

3.11 Bayesian analysis with control function principles

In spite of the apparent success of the classical strategy above, experience suggests Bayesian analysis employing control function principles is more robust than is the classical strategy. Perhaps, this reflects hazard rate (or inverse Mills ratio) sensitivity to estimation error. On the other hand, a Bayesian approach employs least squares estimation on augmented, "complete" data (pseudo-random draws from a truncated normal distribution). That is, instead of extrapolating into the tails via the hazard rate "correction," the Bayesian strategy utilizes data augmentation to "recover" missing counterfactual data.[23]

However, we suspect that it is at least as important that Bayesian analysis helps us or even forces to pay attention to what we know about the setting.[24] Also, Bayesian data augmentation allows the distribution of treatment effects as well as marginal treatment effects to be explored (our discussion above, limits inferences to treatment effect means).[25]

We next turn our attention to Bayesian analysis and consistent reasoning. First, we explore the importance of loss functions, maximum entropy probability assignment, conjugate families, and Bayesian analysis of some

[23] Bayesian analysis is data intensive. Its application to treatment effects is discussed in some detail in *Accounting and Causal Effects: Econometric Challenges*, ch. 12.

[24] Jaynes, 2003, *Probability Theory: The Logic of Science* gives a riveting account of these ideas.

[25] Heckman and others propose classical, factor analytic strategies to explore treatment effect distributions and marginal treatment effects.

primitive data analytic problems. Then, we revisit treatment effects and discuss Bayesian analysis.