working paper

13 10R

Covariates and Causal Effects: The Problem of Context

Dionissi Aliprantis

FEDERAL RESERVE BANK OF CLEVELAND

Working papers of the Federal Reserve Bank of Cleveland are preliminary materials circulated to stimulate discussion and critical comment on research in progress. They may not have been subject to the formal editorial review accorded official Federal Reserve Bank of Cleveland publications. The views stated herein are those of the authors and are not necessarily those of the Federal Reserve Bank of Cleveland or of the Board of Governors of the Federal Reserve System.

Working papers are available on the Cleveland Fed's website at:

www.clevelandfed.org/research.

Working Paper 13-10R August 2014*

Covariates and Causal Effects: The Problem of Context

Dionissi Aliprantis

The literature contrasting structural and experimental econometrics highlights that stronger assumptions are required to identify direct effects than total effects. This paper examines the additional assumptions required to predict future experience using either type of causal effect identified in past data. I show there is a tradeoff between identification and prediction driven by a fact I call the problem of context: Treatment always influences the outcome variable in combination with covariates. The response of covariates to variation in treatment impedes identification of direct effects, while changes over time to the process generating covariates impedes prediction using total effects. I show that total effects can be used for prediction only under restrictive assumptions about the evolution of the DGP: All features of the DGP must remain the same. Direct effects allow for prediction when only some features of the DGP remain the same. To highlight implications for applied work, I discuss the strong assumptions about the behavior of covariates required for prediction when using the total effects of educational attainment successfully identified in past data. While the assumptions required to both identify and predict using standard methodology may be daunting, an explicit statement of these assumptions is a first step towards developing methodology capable of weakening them.

Keywords: Identification, Prediction, Direct Effect, Total Effect, Dynamics, Dynamic Covariates, Dynamic Principal Strata, Lucas Critique, Transportability across Time, Exogeneity.

JEL Codes: C31, C36, C40, C51, C53, I21.

Suggested citation: Aliprantis, Dionissi, 2014. "Covariates and Causal Effects: The Problem of Context," Federal Reserve Bank of Cleveland, working paper no. 13-10R.

Dionissi Aliprantis is at the Federal Reserve Bank of Cleveland; he can be reached at Dionissi.Aliprantis@clev.frb.org. The author thanks Francisca G.-C. Richter, Michela Tincani, Daniel Carroll, Mahmoud Elamin, Martin Huber, Giovanni Mellace, Peter Hinrichs, Nate Baum-Snow, Andrew Butters, Bruce Fallick, James Woodward, and seminar participants at the Cleveland Fed, GATE/Lyon 2, St. Gallen, and the MEA Conference for helpful comments.

*First version August 2013.

1 Introduction

Why are causal effects of greater interest than descriptive statistics? Causal effects are thought to be useful for predicting the future state of the world in ways that descriptive statistics are not.[1] Despite this intended use of causal effects and the careful attention given to the assumptions required for their identification, the assumptions required to predict with causal effects have received little formal treatment.

This paper explicitly connects the identification of causal effects in past data with the prediction of future experience. Focusing on direct and total effects, which are defined in terms of how the corresponding intervention to the Data Generating Process (DGP) affects covariates, I show that while identifying each distinct type of effect requires different assumptions, the same is true for prediction.[2] Unfortunately, the assumptions necessary for identification and prediction do not coincide, and so there is a tradeoff: Stronger assumptions about covariates must be invoked in the identification stage when using direct effects and in the prediction stage when using total effects.[3]

The relative strengths of direct and total effects are generated by a fact I call the problem of context: Treatment always influences the outcome variable in combination with covariates. I show how the relative strength of total effects for identification results from the response of covariates to variation in treatment. This feature of the DGP impedes scientists from identifying direct effects, which requires not only that an intervention would randomize treatment, but that it would also hold covariates at fixed values. In contrast, random variation in treatment is sufficient to identify total effects. The relative difficulty of identifying direct effects has received much attention in the literature (Deaton (2010), Rosenzweig and Wolpin (2000), Keane (2010), White and Chalak (2013), Heckman (1997), Imbens (2010), Freedman (1987), Holland (1988)).

I also show how the relative strength of direct effects for prediction results from changes to the process generating covariates. Remember the problem of induction: Because future experience is outside the support of the data, any prediction is based on unverifiable assumptions about how the DGP evolves over time. A scientist predicting with total effects must make restrictive assumptions

[1] Zellner (2007) follows Jeffreys (2011) and others to distinguish between two key steps of science as being (1) *Description* of the past, and (2) Generalization/*Prediction* of future (or as of yet unobserved) experience. Angrist (2004) notes that "empirical research is almost always motivated by a belief that estimates for a particular context provide useful information about the likely effects of similar programmes or events in the future" (p C52). Similarly, Angrist and Pischke (2009) "believe that the most interesting research in social science is about questions of cause and effect ..., " because "A causal relationship is useful for making predictions about the consequences of changing circumstances or policy; it tells us what would happen in alternative (or 'counterfactual') worlds" (p 3). Heckman and Vytlacil (2007) (pp 4787-92) and Manski (2007) (p 6) contrast this common goal in economics with a competing view that understanding causal effects adds to "knowledge" that is useful in some general sense.

[2] I define covariates as observable (but not necessarily observed) variables other than treatment that causally influence the outcome variable. As discussed in the paper, direct effects are useful for predicting outcomes under interventions allocating treatment while holding the values of covariates fixed (Pearl (2014), Robins et al. (2009)). Total effects predict outcomes under interventions allocating treatment while holding the process generating covariates fixed (Angrist et al. (1996)). Because covariates can respond to treatment as a part of this process, interventions allocating treatment in this way can influence the outcome variable through any number of covariates known as mediators (Imai et al. (2010), VanderWeele (2009), Heckman and Pinto (2013)).

[3] This tradeoff is under-appreciated because internal validity has received more formal attention than external validity in the literature contrasting the structural and experimental approaches devoted to one effect or the other.

about the evolution of the DGP: *All* features of the DGP will remain the same.[4] In contrast, direct effects allow for prediction when only *some* features of the DGP remain the same.

Haavelmo (1944)'s famous example helps to illustrate the identification-prediction tradeoff represented by direct and total effects. Identifying the total effect of gas pedal position on the speed of a car requires only data with random variation in gas pedal position. But in terms of Woodward (2003)'s notion of degrees of invariance, gas pedal position and speed have a weakly invariant relationship.[5] Intervening to change gas pedal position would not change the car's speed in the same way if the process generating any number of covariates were different, such as the incline of the road, fuel type, tire traction, spark plug or overall engine design, overall weight, aerodynamics of the car's body, etc. While the process generating the analogous covariates might not change over time and environment for physical systems, this is not the case in social systems due to human agency. In social systems, the process generating covariates may change enough that the appropriate analogy for the DGP in Haavelmo's example is that of a car frequently altered by a pit crew.

Understanding the identification-prediction tradeoff has major implications for the current literature, which I show by interpreting estimates of returns to schooling in light of the identification-prediction tradeoff. I cite evidence that a long list of covariates that are determined in response to educational attainment also have large effects on wages, including on-the-job training; participation in job training programs; self-employment; vocational education; criminal behavior; arrest; incarceration; fertility; household formation; geographic location; military service; working while in school; smoking; and neighborhood quality. Selection into these covariates is likely to result in violations of the direct effect exclusion restriction requiring randomized and controlled variation in treatment (Deaton (2010), Rosenzweig and Wolpin (2000), Keane (2010)).

In such cases where we do not expect to observe randomized *and* controlled variation in treatment, randomized variation in treatment, and therefore identification of total effects, may be the best we can hope to observe (Imbens (2010)). But thinking in terms of returns to schooling, since human behavior related to selection into the above covariates is likely to change over time, so too are the total effects of the corresponding DGPs likely to change. If total effects of social systems are unstable over time, why should they be more useful than descriptive statistics for predicting the outcome variable under interventions manipulating the treatment variable?

Explicitly connecting identification and prediction links the literature on causal effects to the macroeconomic literature using theory to construct predictions when the process generating co-

[4]When the DGP does not change over time the main obstacle to prediction is linking Local Average Treatment Effects (LATEs, Imbens and Angrist (1994)) relevant to a particular subpopulation to the ATE summarizing information about the DGP for the entire population (Angrist (2004)). Generalizability in terms of subpopulations is only one aspect of external validity, however (Campbell and Stanley (1963), Shadish et al. (2001)). The cross-sectional assumptions that allow a researcher to construct one type of effect from another at a given point in time need not imply the temporal stability assumptions necessary for predicting the future with causal effects (Angrist (2004), Heckman and Vytlacil (2001)). Alternatively, weak exogeneity does not imply super exogeneity (Engle et al. (1983)).

[5]This recent work in philosophy would characterize specific causal effects neither as exceptionless laws of nature nor as complete accidents, but rather as features of a DGP having a degree of invariance located somewhere between these all-or-nothing extremes (Woodward (2000), Woodward (2003)). This analysis has been strongly influenced by Woodward (2003)'s argument that generalizations invariant under some interventions need not be invariant under all others.

variates is dynamic (Lucas (1976), Kydland and Prescott (1977)).[6] If we define credibility as the strength of assumptions required to make a claim (Manski (2007)), then explicitly connecting identification and prediction also highlights that the credibility revolution in empirical economics has been focused on only half of the problem: Identifying causal effects in past data. Because direct and total effects have relative strengths for identification and prediction, so too do the structural and experimental approaches devoted to one effect or the other. From the structural perspective, it is justified to exert extra effort to identify direct effects (Mattei and Mealli (2011), Huber (2013)) and to think about ex-ante evaluation in addition to ex-post evaluation (Wolpin (2007), Carrell et al. (2013)). From the experimental perspective, it is justified to exert extra effort to find identifying variation (Angrist and Pischke (2010), Imbens (2010)) and to investigate whether the empirical content of a structural model is supported by past data (Freedman (1987), Holland (1988)). Finally, explicitly connecting identification and prediction helps to extend the formal analysis of transportability (Only recently begun: See Pearl and Bareinboim (2011), Bareinboim and Pearl (2013b), Bareinboim and Pearl (2013a), Angrist (2004).) to populations for which not even passive observations can be collected (ie, populations in the future).

The paper proceeds as follows: Section 2 defines the set of DGPs to be considered in the paper. I study DGPs in which an outcome variable is causally determined by four variables: treatment, an observed covariate, an unobserved covariate, and an unobservable covariate. I focus on the complications that arise from changes over time to the process generating covariates by considering DGPs in which treatment is both static and randomly assigned at each point in time.[7] Section 3 presents three definitions of causal effects, and this Section, even more than the rest of the paper, owes a strong debt to the lines of research conducted by James Heckman and coauthors and Judea Pearl and coauthors. Section 4 discusses the identification of these causal effects in past data, using four simple DGPs to illustrate its points. This Section shows that distinct exclusion restrictions are required to identify direct and total effects, and it is shown in an Appendix that as a result one person's bias is another person's identification. Section 5 states the assumptions necessary to predict future effects from causal effects identified in past data, and makes clear that researchers' choice between direct effects and total effects represents a tradeoff between the generality of the DGP that can be studied in the past and the generality of the DGP for which the future can be predicted. Section 6 discusses implications for the literature, with a focus on the fact that LATE and MTE estimates of returns to schooling are total effects. Section 7 concludes.

[6] In terms of the analysis in this paper, a simple version of the Lucas (1976) critique can be restated as X_t being an indicator for guards at Fort Knox, D_t being an indicator for an attack on Fort Knox, and Y_t being the gold stolen from Fort Knox with $Y_t \overleftarrow{=} \delta D_t - \delta D_t X_t$. Total effects will change, and therefore so will decisions about whether to intervene to allocate treatment, when all features of DGP stay the same over time except for the process generating the covariate X_t (f_t^X).

[7] This includes, but is not limited to, dynamic decision rules (ie, dynamic principal strata (Frangakis and Rubin (2002)). A more complicated version of the issues studied in this paper also apply to the DGPs studied in the dynamic treatment effects literature where treatment and selection into treatment are dynamic (Robins (1986), Lechner and Miquel (2010), Heckman and Navarro (2007)). The key distinction for the issues studied in this paper is not a focus on selection into covariates over selection into treatment, but rather that the dynamics under consideration will occur from the present moment into the future instead of occurring entirely in the past.

2 Data Generating Processes

Suppose that at time $t \in \mathbb{N}$ data are generated by a Data Generating Process (DGP) in which the outcome variable (Y_{ti}) for each individual i is causally determined by two observed variables, treatment (D_{ti}) and observed covariates (X_{ti}), as well as unmeasured covariates (U^Y_{ti}), or additional factors not observed by the econometrician. The unmeasured covariates can be broken down into those factors that are unobserved (E_{ti}) and those that are unobservable (ϵ_{ti}) at the given level of measurement. The unobserved covariates E_{ti} can be further broken down into permanent (E^P_{ti}) and malleable (E^M_{ti}) factors.

Where variables Z are instruments, U are unmeasured variables, and Θ represents a vector of parameters, the DGP $\mathcal{D}_t$ is characterized by the following structural equations:

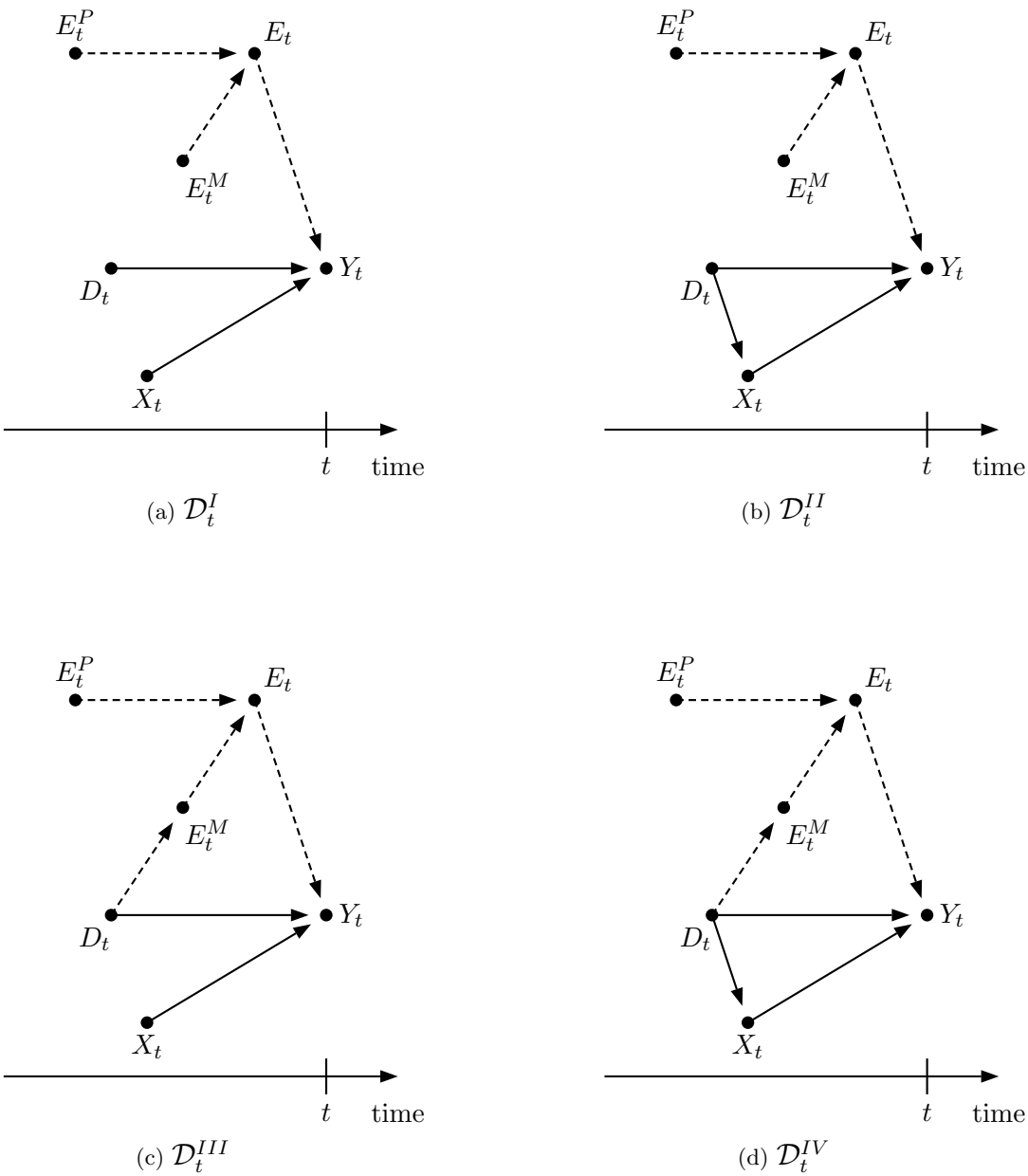
$$
\begin{aligned}
E^P_{ti} &\leftleftarrows f^P_t(U^P_{ti} \ ; \quad \Theta^P_t) && (1)\\
D_{ti} &\leftleftarrows f^D_t(U^D_{ti} \ ; \quad \Theta^D_t) && (2)\\
X_{ti} &\leftleftarrows f^X_t(D_{ti}, U^X_{ti} \ ; \quad \Theta^X_t) && (3)\\
E^M_{ti} &\leftleftarrows f^M_t(D_{ti}, U^M_{ti} \ ; \quad \Theta^M_t) && (4)\\
E_{ti} &\leftleftarrows f^E_t(E^P_{ti}, E^M_{ti} \ ; \quad \Theta^E_t) && (5)\\
Y_{ti} &\leftleftarrows f^Y_t(D_{ti}, X_{ti}, U^Y_{ti} \ ; \quad \Theta_t) \ = \ \theta^0_t + D_{ti}\theta^1_t + X_{ti}\theta^2_t + U^Y_{ti} && (6)\\
&= \ \theta^0_t + D_{ti}\theta^1_t + X_{ti}\theta^2_t + E_{ti}\theta^3_t + \epsilon_{ti}. && (7)
\end{aligned}
$$

The $\leftleftarrows$ notation is formally defined in Section 3, but briefly communicates that the equation contains information about counterfactual manipulations to the right hand side variables, and not only information about what is passively observed in the data. For the sake of exposition we will assume $\theta^3_t = 1$, so that the potential outcome Equation 7 is:

$$Y_{ti} \leftleftarrows \theta^0_t + D_{ti}\theta^1_t + X_{ti}\theta^2_t + E_{ti} + \epsilon_{ti}. \tag{7}$$

An implicit assumption in static models of causal effects is that dependent variables occur a short but finite time interval after independent variables (Pearl (1993), Holland (1986)). To be explicit about this assumption, we assume that at time $t - 6\epsilon \in \mathbb{R}^+$ nature evaluates the arguments of f^P_t in Equation 1, applies f^P_t to them under the parameterization Θ^P_t, and sets the value of E^P_{ti} accordingly, where $0 < \epsilon << 1$. At time $t - 5\epsilon$, nature then proceeds to do the same for Equation 2. Nature proceeds similarly until ultimately finishing at time $t - \epsilon$ by evaluating the arguments of f^Y_t in Equation 7, applying f^Y_t to them under the parameterization Θ_t, and setting the value of Y_{ti} accordingly. The DGP $\mathcal{D}_t$ is indexed by $t \in \mathbb{N}$ because all of the observed random variables in $\mathcal{D}_t$ are observed at time $t \in \mathbb{N}$, after Equations 1-7 have all been evaluated by nature.

Figure 1 shows four DGPs from $\{\mathcal{D}_t\}$, omitting the unmeasured factors in $\mathbf{U}$ except the unobserved factors in U^Y:



(a) $\mathcal{D}_t^I$

(b) $\mathcal{D}_t^{II}$

(c) $\mathcal{D}_t^{III}$

(d) $\mathcal{D}_t^{IV}$

Figure 1: Directed Acyclic Graphs of Four Data Generating Processes from $\{\mathcal{D}_t\}$

Below I present three definitions of causal effects from this set of DGPs so that later issues related to identification and prediction are not confused with issues of definition. Throughout the analysis I distinguish between the model at time t, $\mathcal{M}_t$, which is a set of relationships between variables specified by the econometrician, and the DGP $\mathcal{D}_t$, which is the set of relationships between variables specified by nature (in this case, specified by structural equations 1-7). For the sake of aiding interpretation, the ensuing discussion will consider an example in which we are trying to learn about the effects of precipitation on crop yields. The features of the DGP will often be interpreted as crop yields in year t and field i (Y_{ti}) being determined by rainfall (D_{ti}), irrigated

water (X_{ti}), pesticide (E_{ti}^M), and soil quality as measured by its pH level (E_{ti}^P).

3 Defining Causal Effects as Changes from Interventions to the DGP

Causal effects quantitatively characterize the change in the outcome variable that would result from an intervention to the DGP. Such interventions to the DGP can be characterized by how they would, or would not, impact covariates, especially unmeasured variables. In order to be precise about which features of the DGP are changed, and which are not, under specific interventions, I use Pearl (2009b)'s *do*-operator throughout the remainder of the analysis.

For example, an ideal manipulation is an intervention setting the value of D_{ti} but making no other change to the DGP (Spirtes et al. (2004), Eberhardt and Scheines (2007)). That is, an ideal manipulation changes the process generating D_{ti} but does not change the process generating either covariate X_{ti} or U_{ti}^Y. Such an intervention is written in potential outcome notation as $Y_{ti}(d)$, which can also be communicated in *do*-operator notation as the random variable Y_{ti} after the $do(D_{ti} = d)$ operator has been applied to the DGP.[8]

The $\Leftarrow$ notation in Equations 1-7 is meant to make explicit that these equations convey information about the variation we would observe in the left hand side (*lhs*) variable if we were to intervene to counterfactually vary the right hand side (*rhs*) variables in specific ways (Pearl (2009b)), often specified by the *do*-operator. This counterfactual variation contrasts with the variation we actually do observe in the *lhs* variable as the *rhs* variable varies in the data freely (ie, without intervention).[9] A standard equation conveying information about the relationship between variables observed in the factual world is symmetric with regards to the variables on its left and right hand sides.

A structural equation conveying information about the behavior of the *lhs* variable in counterfactual worlds in which *rhs* variables are under control characterizes an asymmetric relationship between the variables on the left and right hand sides of the equation.[10] For example, structural Equation 7 provides information about the counterfactual values of Y_{ti} if we were to control the variables on the *rhs*, but makes no claim about how any of the variables on the *rhs* would behave if we were to control the *lhs* variable Y_{ti} (Pearl (2009b), p 160). The $\Leftarrow$ notation together with the parameters of the equation makes precise quantitative predictions about the DGP under intervention. In the case of Equation 7, θ_t^1 conveys information about the change in Y_{ti} we would

[8] White and Chalak (2009) and Chalak and White (2012) provide an extension to settable systems, where interventions and causal effects are defined in terms of those causal variables set to particular values by the intervention and those allowed to respond according to their structural equation. Settable systems are especially useful for dealing with background variables and beliefs, and appear to be promising for addressing the issues discussed in this analysis.

[9] Heckman (2005) makes this distinction in terms of setting variables rather than conditioning on them, and in the Rubin Causal Model this distinction is made using potential outcomes notation to separate the causal theory from the observed data (Holland (1988)). Chen and Pearl (2012) discuss some of the confusion that can arise without such explicit notation.

[10] Some examples of the asymmetry of "directions of influence" (Strotz and Wold (1960)) include the fact that symptoms do not cause disease (Pearl (2009a)), a child's height does not cause her father's height (Goldberger (1991)), and rainfall determines crop yields but not the reverse.

counterfactually observe under an intervention varying D_{ti} while *also* keeping covariates X_{ti} and U_{ti}^{Y} constant.

The $\Leftarrow$ notation in Equations 1-7 not only communicates information about how the *lhs* variable in the DGP would respond to interventions to the *rhs* variables, but also how the *lhs* variable would respond to interventions to additional variables. Recall that an exclusion restriction is a feature of the DGP in which the *lhs* variable would not respond under intervention to a set of variables other than those on the *rhs* of the equation. How many variables other than those on the *rhs* must admit to an exclusion restriction in order for an equation to be structural?

Throughout this paper I use the symbol $\Leftarrow$ to denote that an equation is structural in the sense that given control over the specified observable (but not necessarily observed) variables on the *rhs* of the equation, changes made to additional observable variables would provide no further change to the outcome variable. This follows a standard definition of "structural equation" requiring that all variables at a given level of observation not included on the right hand side satisfy an exclusion restriction (Pearl (2009b), Definition 5.4.1). However, an alternative definition of "structural equation" requires that an exclusion restriction must hold for only one additional variable not included on the *rhs* (Angrist et al. (1996)). It is important to distinguish between these definitions because the same intervention applied to two DGPs identical in all ways except for their definition of "structural equation" could result in distinct counterfactual outcomes.

3.1 The Direct Effect: Changes from Interventions to Treatment and Covariates

Direct effects characterize the change in the outcome variable from a specific type of intervention to the DGP:

Definition D1 (The Direct Effect): The individual-level causal effect of D_{ti} on Y_{ti}, Δ_{ti}^{D1}, is the change in Y_{ti} that would result from an intervention setting D_{ti} to various values while setting all other variables entering as arguments in f^{Y} to fixed values.

Direct effects are typically defined relative to a reference set of variables that are set to fixed values by an intervention, and the values to which the reference variables are set (See Pearl (2001) for an introduction.). In general, direct effects need not be invariant to interventions changing variables outside that reference set. This direct effect, however, is invariant to changes to variables outside the given reference set. This is because D1 defines direct effect relative to the reference set that is the variables on the *rhs* of Equation 7, and the definition of structural equation adopted in this paper indicates that *all* variables at the given level of observation not included on the *rhs* of Equation 7 satisfy an exclusion restriction. In other words, because we are examining the DGP and not a model, mediators outside the reference set can only be found at a finer level of observation (A process which Holland (1988) notes always appears to be possible.).

The fundamental problem of causal inference is that we can only observe individuals in one state of treatment at a given point in time (Holland (1986)). Thus causal inference, even entirely in the past, is impossible without constructing unobserved counterfactuals. Since analysts are

typically more comfortable constructing the average counterfactual for some population rather than individual-level counterfactuals, it is useful to rewrite the individual-level effect in D1 in terms of the Average Causal Effect (ACE) using the *do* operator. Equation 7 and definition D1 could be interpreted as indicating that

Definition D1 (The Direct ACE): The Average Causal Effect of D_t on Y_t is

$$\begin{aligned}\mathbb{E}_i[\Delta_{ti}^{D1}(d,d')] &\equiv \mathbb{E}_i[Y_{ti}|do(D_{ti}=d', X_{ti}=x_{ti}, U_{ti}^Y=u_{ti})] \;-\; \mathbb{E}_i[Y_{ti}|do(D_{ti}=d, X_{ti}=x_{ti}, U_{ti}^Y=u_{ti})] \\ &= \theta_t^1(d'-d).\end{aligned}$$

3.2 The Total Effect: Changes from Interventions to Treatment Alone

A second useful definition of causal effect is the one adopted in the Rubin Causal Model:

Definition D2 (The Total Effect): The individual-level causal effect of D_{ti} on Y_{ti}, Δ_{ti}^{D2}, is the change in Y_{ti} that would result from an intervention setting D_{ti} to various values without changing any other feature of the DGP.

The analogous definition in Angrist et al. (1996) also requires that this effect is observable through the existence of an intervention generating $Y_{ti}(d)$ (ie, an intervention that sets D_{ti} while not affecting the processes generating any of the covariates influencing Y_{ti}). Such an intervention corresponds to the existence of an instrumental variable Z_{ti}^D that enters as an argument of f_t^D in Equation 2 but is excluded from being an argument in all other equations. This exclusion restriction for total effects allows for D_{ti} to be an argument in the selection Equations 3 and 4. Thus we might write

$$Y_{ti}(D_{ti}) \leftleftarrows \theta_t^0 + D_{ti}\theta_t^1 + X_{ti}(D_{ti})\theta_t^2 + U_{ti}^Y(D_{ti}), \tag{1*}$$

so that under definition D2 the individual-level causal effect of D_{ti} on Y_{ti} can be decomposed into the effect of D_{ti} on Y_{ti} from definition D1 together with the change in Y_{ti} that would result from the change in covariates induced by the change in D_{ti}:

$$\begin{aligned}\Delta_{ti}^{D2}(d,d') \equiv Y_{ti}(d') - Y_{ti}(d) &= \left[\theta_t^0 + d'\theta_t^1 + X_{ti}(d')\theta_t^2 + U_{ti}^Y(d')\right] \;-\; \left[\theta_t^0 + d\theta_t^1 + X_{ti}(d)\theta_t^2 + U_{ti}^Y(d)\right] \\ &= \theta_t^1 \left[\; d' - d \;\right] \;+\; \theta_t^2 \left[\; X_{ti}(d') - X_{ti}(d) \;\right] \;+\; \left[\; U_{ti}^Y(d') - U_{ti}^Y(d) \;\right].\end{aligned}$$

We might again define causal effects in terms of the ACE in response to the fundamental problem of evaluation:

Definition D2 (The Total ACE): The Average Causal Effect of D_t on Y_t is

$$\mathbb{E}_i[\Delta_{ti}^{D2}(d,d')] \equiv \mathbb{E}_i[Y_{ti}|do(D_{ti}=d')] - \mathbb{E}_i[Y_{ti}|do(D_{ti}=d)].$$

3.3 The Econometric/Graphical Definition

Defining the vector

$$S \equiv (D, X, U^Y),$$

we can re-write Equation 7 more compactly as

$$Y_{ti} \leftleftarrows Y_{ti}(S_{ti}).$$

Heckman (2008) and Pearl (2009b) (Definition 3.2.1) define the individual-level causal effect as

D3 The causal effect of manipulating s to s' is $\Delta_{ti}^{D3}(s, s') \equiv Y_{ti}(s') - Y_{ti}(s)$ when $s \neq s'$.

Both the direct and total definitions of causal effect are nested within this definition. Let $s = (d, x, u^Y)$. If $s' = (d', x, u^Y)$, then the causal effect in definition D3 corresponds to the direct definition. If $s' = (d', x(d'), u^Y(d'))$, then the causal effect in D3 corresponds to the total definition. Definitions of causal effects in addition to direct and total effects correspond to other possible settings of or rules for determining s' (Pearl (2014)).

3.4 Representing Interventions to the DGP

Both direct and total effects are defined in terms of the interventions to the DGP they characterize. An exclusion restriction is necessary for defining the intervention resulting in total effects (Angrist et al. (1996)), and an inclusion restriction is additionally required for defining direct effects. Let Z_{ti}^T represent an intervention resulting in a change in Y_{ti} described by the total effect, and Z_{ti}^D represent an intervention resulting in a change in Y_{ti} described by the direct effect. Then the exclusion restriction requires that Z_{ti}^T has no influence on the process generating covariates (or parents other than D_{ti}) of Y_{ti}. The inclusion restriction requires that Z_{ti}^D not only influences the process generating covariates of Y_{ti}, but completely controls it.

These interventions represent specific changes to the DGPs described by Equations 1-7:

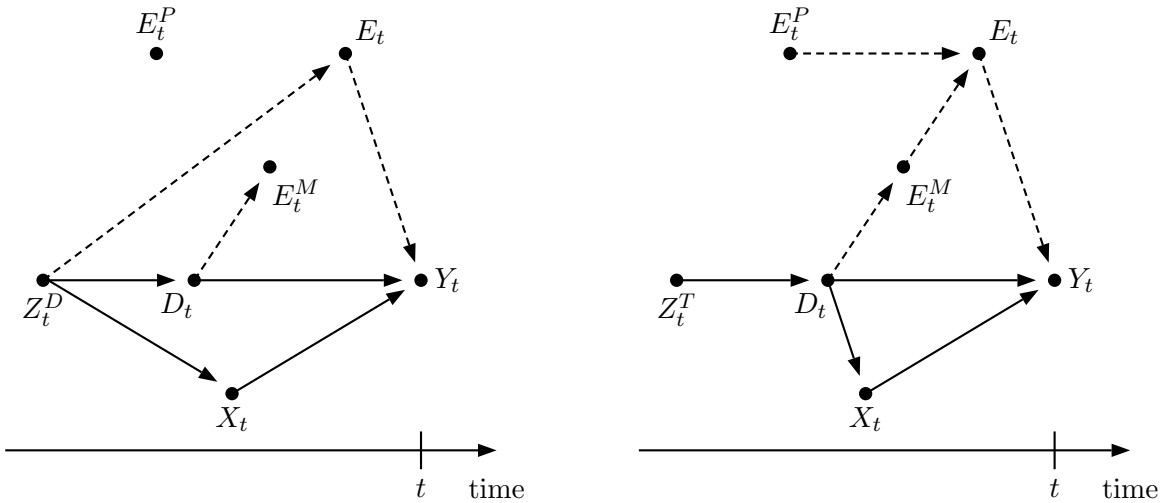
Original DGP	After Direct Intervention	After Total Intervention
$E_{ti}^P \leftleftarrows f_t^P(U_{ti}^P \ ; \Theta_t^P)$	$E_{ti}^P \leftleftarrows f_t^P(U_{ti}^P \ ; \Theta_t^P)$	$E_{ti}^P \leftleftarrows f_t^P(U_{ti}^P \ ; \Theta_t^P)$
$D_{ti} \leftleftarrows f_t^D(U_{ti}^D \ ; \Theta_t^D)$	$D_{ti} \leftleftarrows f_t^D(\mathbf{Z_{ti}^D}, U_{ti}^D \ ; \Theta_t^D)$	$D_{ti} \leftleftarrows f_t^D(\mathbf{Z_{ti}^T}, U_{ti}^D \ ; \Theta_t^D)$
$X_{ti} \leftleftarrows f_t^X(D_{ti}, U_{ti}^X \ ; \Theta_t^X)$	$X_{ti} \leftleftarrows f_t^X(\mathbf{Z_{ti}^D}, D_{ti}, U_{ti}^X \ ; \Theta_t^X)$	$X_{ti} \leftleftarrows f_t^X(D_{ti}, U_{ti}^X \ ; \Theta_t^X)$
$E_{ti}^M \leftleftarrows f_t^M(D_{ti}, U_{ti}^M \ ; \Theta_t^M)$	$E_{ti}^M \leftleftarrows f_t^M(D_{ti}, U_{ti}^M \ ; \Theta_t^M)$	$E_{ti}^M \leftleftarrows f_t^M(D_{ti}, U_{ti}^M \ ; \Theta_t^M)$
$E_{ti} \leftleftarrows f_t^E(E_{ti}^P, E_{ti}^M \ ; \Theta_t^E)$	$E_{ti} \leftleftarrows f_t^E(\mathbf{Z_{ti}^D}, E_{ti}^P, E_{ti}^M \ ; \Theta_t^E)$	$E_{ti} \leftleftarrows f_t^E(E_{ti}^P, E_{ti}^M \ ; \Theta_t^E)$
$Y_{ti} \leftleftarrows f_t^Y(D_{ti}, X_{ti}, U_{ti}^Y \ ; \ \Theta_t)$	$Y_{ti} \leftleftarrows f_t^Y(D_{ti}, X_{ti}, U_{ti}^Y \ ; \ \Theta_t)$	$Y_{ti} \leftleftarrows f_t^Y(D_{ti}, X_{ti}, U_{ti}^Y \ ; \ \Theta_t)$

resulting in the altered DGPs corresponding with the definition of causal effect in Pearl (2009b)

(Definition 3.2.1):

Original DGP	After Direct Intervention	After Total Intervention
$E_{ti}^P \leftleftarrows f_t^P(U_{ti}^P \ ;\Theta_t^P)$	$E_{ti}^P \leftleftarrows f_t^P(U_{ti}^P \ ;\Theta_t^P)$	$E_{ti}^P \leftleftarrows f_t^P(U_{ti}^P \ ;\Theta_t^P)$
$D_{ti} \leftleftarrows f_t^D(U_{ti}^D \ ;\Theta_t^D)$	$D_{ti} \leftleftarrows d$	$D_{ti} \leftleftarrows d$
$X_{ti} \leftleftarrows f_t^X(D_{ti}, U_{ti}^X \ ;\Theta_t^X)$	$X_{ti} \leftleftarrows x$	$X_{ti} \leftleftarrows f_t^X(D_{ti}, U_{ti}^X \ ;\Theta_t^X)$
$E_{ti}^M \leftleftarrows f_t^M(D_{ti}, U_{ti}^M \ ;\Theta_t^M)$	$E_{ti}^M \leftleftarrows f_t^M(D_{ti}, U_{ti}^M \ ;\Theta_t^M)$	$E_{ti}^M \leftleftarrows f_t^M(D_{ti}, U_{ti}^M \ ;\Theta_t^M)$
$E_{ti} \leftleftarrows f_t^E(E_{ti}^P, E_{ti}^M \ ;\Theta_t^E)$	$E_{ti} \leftleftarrows e$	$E_{ti} \leftleftarrows f_t^E(E_{ti}^P, E_{ti}^M \ ;\Theta_t^E)$
$Y_{ti} \leftleftarrows f_t^Y(D_{ti}, X_{ti}, U_{ti}^Y \ ; \ \Theta_t)$	$Y_{ti} \leftleftarrows f_t^Y(D_{ti}, X_{ti}, U_{ti}^Y \ ; \ \Theta_t)$	$Y_{ti} \leftleftarrows f_t^Y(D_{ti}, X_{ti}, U_{ti}^Y \ ; \ \Theta_t)$

Figure 2 shows the new DGP after interventions Z_{ti}^D or Z_{ti}^T are applied to DGP $\mathcal{D}_t^{IV}$:



(a) Intervention Z_t^D to $\mathcal{D}_t^{IV}$ Described by Direct Effect (b) Intervention Z_t^T to $\mathcal{D}_t^{IV}$ Described by Total Effect

Figure 2: Directed Acyclic Graphs of Two Interventions to the DGP $\mathcal{D}_t^{IV}$

An implicit assumption about both the direct and total interventions is that they do not change the process generating the unobservable component ϵ_{ti} of the unmeasured covariates U_{ti}^Y. This is a crucial assumption. Since defining a feature of a DGP and observing that feature are separate tasks, the inability to observe a defined feature does not negate the operationality of the definition (See p 160 of Pearl (2009b) or Heckman and Vytlacil (2007).). But it does appear difficult for a researcher to even theorize about an intervention holding all else constant (Freedman (1987), Holland (1988), Rogosa (1987)) or not influencing all else in any way (Eberhardt and Scheines (2007)) without knowing what variables are included in "all else."[11]

[11] Some of the criticisms in the cited papers could also be seen as arguments that when economists "structurally estimate" a model, meaning that they estimate *a model* $\mathcal{M}_t$ whose equations are assumed to be invariant in some way, it appears incredibly difficult to judge whether such invariance conditions characterize the real world (ie, *the DGP* $\mathcal{D}_t$). The use of non-random holdout samples is a promising approach for judging whether the invariance hypothesized by a researcher's model $\mathcal{M}_t$ corresponds with that of the DGP $\mathcal{D}_t$ (Todd and Wolpin (2006), Keane and Wolpin (2007), Schorfheide and Wolpin (2013)).

Table 1 describes several interventions to the DGP in terms of both potential outcome notation (ie, the RCM) and structural equations notation (ie, the SCM). The recent literature has studied the identification of a plethora of such causal effects, where each particular causal effect is defined in terms of which covariates the intervention holds fixed or allows to respond to other variables in the DGP (Pearl (2001), Pearl (2014), Hafeman and VanderWeele (2011), Huber (2013), Flores and Flores-Lagunes (2010)).

Table 1: Alternative Notation for Identical Interventions to the DGP $\mathcal{D}_t$

Directed Acyclic Graphs/ Structural Causal Model	Econometric Model	Potential Outcomes/ Rubin Causal Model	IV/Regression Coefficients Under Appropriate Assumptions
$\mathbb{E}[Y_t \vert D_t = d]$	$\mathbb{E}[Y_t \vert S_t = s]$	$\mathbb{E}[Y_t \vert D_t = d]$	$\beta_t d + \epsilon$
$\mathbb{E}[Y_t \vert do(D_t = d')]-$ $\mathbb{E}[Y_t \vert do(D_t = d)]$	$\mathbb{E}[Y_t(s')]-$ $\mathbb{E}[Y_t(s)]$	$\mathbb{E}[Y_t(d')]-$ $\mathbb{E}[Y_t(d)]$	$\beta_t(d' - d)$ in RCM
$\mathbb{E}[Y_t \vert do(D_t = d', X_t = m_x)]-$ $\mathbb{E}[Y_t \vert do(D_t = d, X_t = m_x)]$	$\mathbb{E}[Y_t(s')]-$ $\mathbb{E}[Y_t(s)]$	$\mathbb{E}[Y_t(d', m_x)]-$ $\mathbb{E}[Y_t(d, m_x)]$	
$\mathbb{E}[Y_t \vert do(D_t = d', X_t = f_t^X(d, U_t^X; \Theta_t^X))]-$ $\mathbb{E}[Y_t \vert do(D_t = d, X_t = f_t^X(d, U_t^X; \Theta_t^X))]$	$\mathbb{E}[Y_t(s')]-$ $\mathbb{E}[Y_t(s)]$	$\mathbb{E}[Y_t(d', m_{xt}(d))]-$ $\mathbb{E}[Y_t(d, m_{xt}(d))]$	
$\mathbb{E}[Y_t \vert do(D_t = d', X_t = m_x, U_t^Y = m_u)]-$ $\mathbb{E}[Y_t \vert do(D_t = d, X_t = m_x, U_t^Y = m_u)]$	$\mathbb{E}[Y_t(s')]-$ $\mathbb{E}[Y_t(s)]$	$\mathbb{E}[Y_t(d', m_x, m_u)]-$ $\mathbb{E}[Y_t(d, m_x, m_u)]$	$\beta_t(d' - d)$ in SCM

4 Description of the Past: Identification of Causal Effects

To facilitate interpretation, the ensuing analysis will at times discuss identification of causal effects of the DGP specified in the structural Equations 1-7 in terms of the causal effects of precipitation.[12] Suppose first that the researcher is interested in understanding the effects of precipitation on crop yields, and has data on crop yields for farmers $i = 1, \ldots, N$ (Y_{ti} in Equation 7) at times t_0 and t, with $\boldsymbol{Y} \in \mathbb{R}^N \times \mathbb{R}^N$. The researcher might be interested in identifying the effect of precipitation (D_{ti}), which is observed, on crop yields. The social scientist observes another causal variable in the DGP, which is a measure of the quantity of irrigated water applied to the crops (X_{ti}), but does not measure several other causal variables, such as the quantity of pesticide applied to the crops (potentially observable at the given level of measurement) or the quality of the soil in which the crops were grown (assumed to be unobservable at the given level of measurement). The total effect of these unobserved causal variables is labeled as U_{ti}^Y. To re-iterate, we will denote the current moment in time $t^* \in \mathbb{N}$, and we will consider DGPs indexed to three additional points in time:

$$t_0 < t < t^* < t'.$$

[12] Deuchert and Huber (2014) present a related discussion of timing and identification in a setting of causal effects of educational attainment and veteran status on health.

We interpret t_0 as the time at which treatment was assigned, t as the time at which observed variables were measured, and t' being some point in time in the future.

Given the dearth of attention paid to selection into covariates, here we assume selection into treatment is not a problem by assuming that treatment is randomly assigned:

Random Assignment (RA): The variable D_{ti} is randomly allocated in the sample. Specifically, D_{ti} is an *iid* random variable that follows the triangle distribution with lower limit -1, upper limit 1, and mode 0.

Thinking in terms of our rainfall and crop yield example, we can think of this assumption applying to circumstances in which annual rainfall is randomly allocated.

In this scenario a multitude of identification strategies have been developed to identify a variety of causal effects in addition to those defined in Section 3, a short group of which includes Pearl (2001), Chalak and White (2011), and Huber (2013). In this analysis we consider whether the parameters from three regression equations estimated via OLS identify any of the causal effects defined in Section 3, where H, K, and L are statistical error terms:[13]

$$Y_{ti} = \alpha_t^0 + D_{ti}\alpha_t^1 + H_{ti}, \tag{8}$$

$$Y_{ti} = \beta_t^0 + D_{ti}\beta_t^1 + X_{ti}\beta_t^2 + K_{ti}, \tag{9}$$

$$Y_{ti} = \gamma_t^0 + D_{ti}\gamma_t^1 + X_{t_0 i}\gamma_t^2 + L_{ti}. \tag{10}$$

Equation 9 can be justified as the classical regression model with a few additional assumptions (Goldberger (1991), Chapters 15 and 16). And when it is suspected that D_{ti} is an argument in selection Equation 3, Angrist and Pischke (2009) recommend estimating Equation 8 (p 66) and Duflo et al. (2007) (p 3949) and Rosenbaum (1984) recommend estimating Equation 10.

In Appendix A it is shown that when (D_{ti}, X_{ti}, E_{ti}) are excluded from f_{ti}^{ϵ},

$$\widehat{\alpha}_t^{1,OLS} \xrightarrow{p} \theta_t^1 + \frac{\mathbb{E}\left[D_t X_t\right]}{\mathbb{E}\left[D_t D_t\right]}\theta_t^2 + \frac{\mathbb{E}\left[D_t E_t\right]}{\mathbb{E}\left[D_t D_t\right]}, \tag{11}$$

$$\widehat{\beta}_t^{1,OLS} \xrightarrow{p} \theta_t^1 + \frac{\mathbb{E}[X_t X_t]\,\mathbb{E}[D_t E_t] \;-\; \mathbb{E}[D_t X_t]\,\mathbb{E}[X_t E_t]}{\mathbb{E}[D_t D_t]\,\mathbb{E}[X_t X_t] \;-\; \mathbb{E}[D_t X_t]\,\mathbb{E}[X_t D_t]}, \quad \text{and} \tag{12}$$

$$\begin{aligned}\widehat{\gamma}_t^{1,OLS} \xrightarrow{p} \theta_t^1 &+ \frac{\mathbb{E}[X_{t_0} X_{t_0}]\,\mathbb{E}[D_t X_t] \;-\; \mathbb{E}[D_t X_{t_0}]\,\mathbb{E}[X_{t_0} X_t]}{\mathbb{E}[D_t D_t]\,\mathbb{E}[X_{t_0} X_{t_0}] \;-\; \mathbb{E}[D_t X_{t_0}]\,\mathbb{E}[X_{t_0} D_t]}\theta_t^2 \\ &+ \frac{\mathbb{E}[X_{t_0} X_{t_0}]\,\mathbb{E}[D_t E_t] \;-\; \mathbb{E}[D_t X_{t_0}]\,\mathbb{E}[X_{t_0} E_t]}{\mathbb{E}[D_t D_t]\,\mathbb{E}[X_{t_0} X_{t_0}] \;-\; \mathbb{E}[D_t X_{t_0}]\,\mathbb{E}[X_{t_0} D_t]}.\end{aligned} \tag{13}$$

[13] I focus on OLS estimators applied to a randomized treatment because I want to focus on the issues raised by selection into covariates in isolation from those raised by selection into treatment. I show in Appendix A that the OLS estimators of α_t^1, β_t^1, and γ_t^1 converge in probability to the analogous 2SLS estimators when there is perfect compliance between the instrument and treatment. I also provide simulation results in Appendix C showing that the analysis generalizes immediately to identification schemes using instrumental variables estimators to overcome the distinct issue of selection into treatment.

Since D_{ti} has mean zero, the following orthogonality conditions result from **RA**:

Orthogonality DX-t_0: $\mathbb{E}[D_t X_{t_0}] = 0$

Orthogonality DE-t_0: $\mathbb{E}[D_t E_{t_0}] = 0$

Remarkably, one of the orthogonality conditions induced by randomization (**DX-t_0**) implies that $\widehat{\alpha}_t^{1,OLS} \xrightarrow{p} \widehat{\gamma}_t^{1,OLS}$ as $N \to \infty$. Thus Equations 11-13 can be rewritten as:

$$\widehat{\alpha}_t^{1,OLS} \xrightarrow{p} \widehat{\gamma}_t^{1,OLS} \xrightarrow{p} \theta_t^1 + \frac{\mathbb{E}\left[D_t X_t\right]}{\mathbb{E}\left[D_t D_t\right]}\theta_t^2 + \frac{\mathbb{E}\left[D_t E_t\right]}{\mathbb{E}\left[D_t D_t\right]}, \quad \text{and} \tag{14}$$

$$\widehat{\beta}_t^{1,OLS} \xrightarrow{p} \theta_t^1 + \frac{\mathbb{E}[X_t X_t]\,\mathbb{E}[D_t E_t] \;-\; \mathbb{E}[D_t X_t]\,\mathbb{E}[X_t E_t]}{\mathbb{E}[D_t D_t]\,\mathbb{E}[X_t X_t] \;-\; \mathbb{E}[D_t X_t]\,\mathbb{E}[X_t D_t]}. \tag{15}$$

Not only does $\alpha_t^{1,OLS}$ convey the change in the outcome Y_{ti} for each unit of change in D_{ti} we observe in the data,

$$\begin{aligned}
\mathbb{E}[Y_{ti}|D_{ti} = d+1] - \mathbb{E}[Y_{ti}|D_{ti} = d] &= \mathbb{E}[\theta_t^0 + D_{ti}\theta_t^1 + X_{ti}\theta_t^3 + E_{ti} \,|\, D_{ti} = d+1] \\
&\quad - \mathbb{E}[\theta_t^0 + D_{ti}\theta_t^1 + X_{ti}\theta_t^2 + E_{ti} \,|\, D_{ti} = d] \\
&= (d+1)\theta_t^1 + \mathbb{E}[X_{ti}\theta_t^2|D_{ti} = d+1] + \mathbb{E}[E_{ti} \,|\, D_{ti} = d+1] \qquad (16) \\
&\quad - d\theta_t^1 - \mathbb{E}[X_{ti}\theta_t^2|D_{ti} = d] + \mathbb{E}[E_{ti}|D_{ti} = d] \\
&= \theta_t^1 + \frac{\mathbb{E}\left[D_t X_t\right]}{\mathbb{E}\left[D_t D_t\right]}\theta_t^2 + \frac{\mathbb{E}\left[D_t E_t\right]}{\mathbb{E}\left[D_t D_t\right]} \qquad (17) \\
&= plim\ \widehat{\alpha}_t^{1,OLS},
\end{aligned}$$

but Random Assignment (**RA**) implies that it also identifies the change in Y_{ti} that would result if we were to counterfactually set treatment D_{ti} one unit higher. Specifically, **RA** implies that the average covariate for the subpopulation with D_{ti} observed to be d is equal to the average covariate when D_{ti} is set to d for the population:[14]

$$\mathbb{E}[E_{ti}|D_{ti} = d] = \mathbb{E}[E_{ti}|do(D_{ti} = d)], \text{ and} \tag{18}$$

$$\mathbb{E}[X_{ti}|D_{ti} = d] = \mathbb{E}[X_{ti}|do(D_{ti} = d)]. \tag{19}$$

Equations 18 and 19 are a restatement of the independence assumption in Holland (1986) using the

[14] Equations 18 and 19 do not ensure that the treated and nontreated groups are equal in all aspects apart from the treatment status (Heckman (1996)). The related assumptions in the program evaluation hold because E_{ti} in Equation 7 is a different random variable than the error terms in the program evaluation literature (Imbens and Wooldridge (2009), Blundell and Dias (2009), Angrist et al. (1996)). E_{ti} in Equation 7 is also a different random variable than the error terms in the Conditional Expectation Functions (CEFs) discussed in Goldberger (1991) and Angrist and Pischke (2009).

do operator, and are the link by which conditioning on treatment status identifies the total effect:

$$\begin{aligned}
\mathbb{E}[Y_{ti}|D_{ti}=d+1]-\mathbb{E}[Y_{ti}|D_{ti}=d] &= \mathbb{E}[\theta_t^0+D_{ti}\theta_t^1+X_{ti}\theta_t^3+E_{ti}\,|\,D_{ti}=d+1] \\
&\quad -\mathbb{E}[\theta_t^0+D_{ti}\theta_t^1+X_{ti}\theta_t^2+E_{ti}\,|\,D_{ti}=d] \\
&= (d+1)\theta_t^1+\mathbb{E}[X_{ti}\theta_t^2|D_{ti}=d+1]+\mathbb{E}[E_{ti}\,|\,D_{ti}=d+1] \\
&\quad -d\theta_t^1-\mathbb{E}[X_{ti}\theta_t^2|D_{ti}=d]-\mathbb{E}[E_{ti}|D_{ti}=d] \\
&= \theta_t^1+\mathbb{E}[X_{ti}\theta_t^2|do(D_{ti}=d+1)]+\mathbb{E}[E_{ti}|do(D_{ti}=d+1)] \\
&\quad -\mathbb{E}[X_{ti}\theta_t^2|do(D_{ti}=d)]-\mathbb{E}[E_{ti}|do(D_{ti}=d)] \\
&= \mathbb{E}[Y_{ti}|do(D_{ti}=d+1)]-\mathbb{E}[Y_{ti}|do(D_{ti}=0)]. \qquad (20)
\end{aligned}$$

Since D_{ti} is randomly assigned $\mathbb{E}[Y_{ti}|D_{ti}=d]=\mathbb{E}[Y_{ti}|D_{ti}=d,\,X_{t_0i}=x]$, so that the preceding arguments also show that

$$plim\ \widehat{\gamma}_t^{1,OLS}=plim\ \widehat{\alpha}_t^{1,OLS}=\mathbb{E}_i[Y_{ti}|do(D_{ti}=d+1)]-\mathbb{E}_i[Y_{ti}|do(D_{ti}=d)]\equiv\mathbb{E}_i[\Delta_{ti}^{D2}(d,d+1)].$$

Thus $\widehat{\alpha}_t^{1,OLS}$ and $\widehat{\gamma}_t^{1,OLS}$ identify the total effect of treatment on the outcome variable for our DGP.

In contrast, one can immediately see from Equations 14 and 15 that the orthogonality conditions **DX-t_0** and **DE-t_0** resulting from randomization are not sufficient for any of the regression coefficients in Equations 8-10 to identify the direct effect of treatment θ_t^1. This point has received considerable attention in the literature (Deaton (2010), Rosenzweig and Wolpin (2000), Keane (2010), White and Chalak (2013), Heckman (1997), Leamer (2010), Heckman and Smith (1995)).

Consider additional orthogonality conditions at the time of measurement:

Orthogonality DX-t $\mathbb{E}[D_tX_t]=0$

Orthogonality DE-t $\mathbb{E}[D_tE_t]=0$

Orthogonality XE-t $\mathbb{E}[X_tE_t]=0$.

If DE-t and XE-t hold, then $\widehat{\beta}_t^{1,OLS}$ will converge in probability to the direct effect θ_t^1. Note again, however, that randomization is not sufficient to ensure these orthogonality conditions, as it is possible that unobserved covariates E_{ti} are chosen in response either to treatment or observed covariates, which may themselves be chosen in response to treatment.

4.1 The Identification Tradeoff Favoring Total Effects

Recalling that $t_0<t<t^*<t'\in\mathbb{N}$, let $\mathcal{I}_{t^*}(\Delta_t)$ be the subset of DGPs in $\{\mathcal{D}_t\}$ for which causal effects of type Δ_t are identified at the present time t^* by one of the OLS estimators.[15] The preceding analysis showed that total effects are identified by one of the OLS estimators for all

[15] I again stress that there are many other identification strategies for identifying various effects like the direct and total effects as defined in this paper, as well as others like the natural direct effect. See Chalak and White (2011), Flores and Flores-Lagunes (2010), Mattei and Mealli (2011), Huber (2013), and others.

DGPs, but that an OLS estimator can only identify direct effects when orthogonality conditions DE-t and XE-t hold. Given the structure of our DGPs, this happens when neither D_{ti} nor X_{ti} enter as arguments in f_t^M or f_t^E, and E_{ti}^P does not enter as an argument in f_t^X. That is:

$$\mathcal{I}_{t^*}(\Delta_t^{D1}) = \left\{ \mathcal{D}_t \;\middle|\; \begin{array}{l} E_{ti}^P \notin f_t^X, \\ D_{ti}, \notin f_t^M, \quad X_{ti} \notin f_t^M, \\ D_{ti} \notin f_t^E, \quad X_{ti} \notin f_t^E \end{array} \right\}, \text{ while}$$

$$\mathcal{I}_{t^*}(\Delta_t^{D2}) = \left\{ \mathcal{D}_t \right\}.$$

Since any element $\mathcal{D}_t^{D1} \in \mathcal{I}_{t^*}(\Delta_t^{D1})$ is by definition also an element of $\mathcal{I}_{t^*}(\Delta_t^{D2})$, it follows that $\mathcal{I}_{t^*}(\Delta_t^{D1}) \subseteq \mathcal{I}_{t^*}(\Delta_t^{D2})$. In contrast, there exist DGPs $\mathcal{D}_t^{D2} \in \mathcal{I}_{t^*}(\Delta_t^{D2})$ for which D_{ti} is an argument in f_t^M, such as the DGPs $\mathcal{D}_t^{III}$ or $\mathcal{D}_t^{IV}$ specified in Section 9. Since these DGPs violate Exclusion II, $\mathcal{D}_t^{III} \notin \mathcal{I}_{t^*}(\Delta_t^{D1})$ and $\mathcal{D}_t^{IV} \notin \mathcal{I}_{t^*}(\Delta_t^{D1})$, and hence $\mathcal{D}_t^{D2} \not\subset \mathcal{I}_{t^*}(\Delta_t^{D1})$. Thus we have the identification tradeoff favoring total effects: $\mathcal{I}_{t^*}(\Delta_t^{D1})$ is a proper subset of $\mathcal{I}_{t^*}(\Delta_t^{D2})$, or

$$\mathcal{I}_{t^*}(\Delta_t^{D1}) \subset \mathcal{I}_{t^*}(\Delta_t^{D2}).$$

5 Generalization: Prediction with Causal Effects

Suppose that a social scientist has successfully identified a causal effect of the DGP $\mathcal{D}_t$. Why would anyone be interested in such information? If we follow Zellner (2007) to distinguish between two key steps of science being (1) *Description* of the past, and (2) Generalization/*Prediction* of future (or as of yet unobserved) experience, social scientists' interest in causal effects is typically justified in terms of their use for *Prediction*.[16]

How might a causal effect identified in past data be used to predict future experience? We again consider this question in terms of our running example of rainfall (D) and crop yields (Y). Figure 3 shows an example DGP determining Y at time t ($\mathcal{D}_t$), along with its successor DGP determining Y at time $t' > t$ ($\mathcal{D}_{t'}$). If a scientist at the current moment in time (t^*) knows the value of a causal effect identified from data in the past (at time t), what assumptions must hold in order for her to accurately predict how the outcome variable will change if an intervention manipulates treatment in the future (at time t')?

[16] Zellner himself follows Karl Pearson, Harold Jeffreys, and others in this distinction. See Footnote 1 for prominent researchers citing *Prediction* as a justification for interest in causal effects.

A change in the environmental regulation of pesticide use between times t and t' and the resulting realizations of $\mathcal{D}_t$ and $\mathcal{D}_{t'}$

E^P = Soil Quality
E^M = Pesticide
X = Irrigated Water
D = Rainfall
Y = Crop Yield

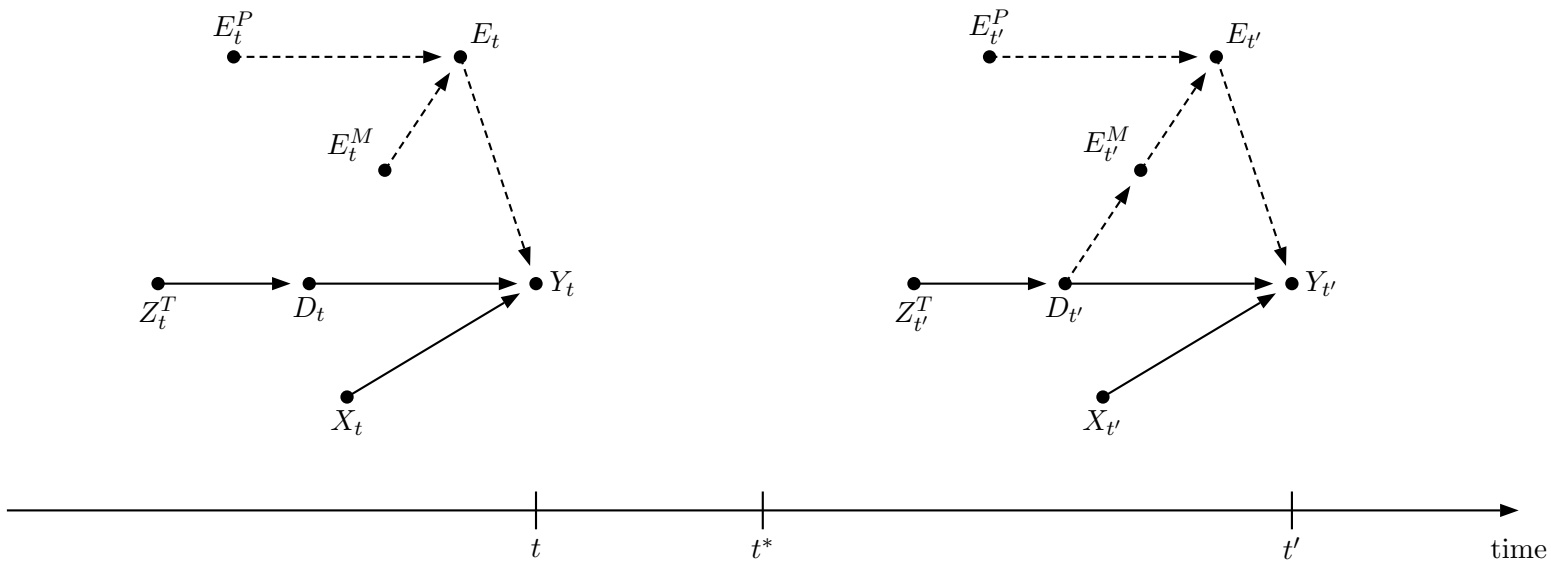


Figure 3: An example in which predictions of the future effect of $D_{t'}$ on $Y_{t'}$ (ie, of changes to $Y_{t'}$ from the intervention $Z_{t'}^T$) will be biased when constructed at current time t^* using total effects identified from past data (ie, from data collected at time t)

5.1 Prediction with Direct Effects

Letting $\Pi_s(\cdot|t)$ denote scientist s's prediction at time t, suppose that the following three assumptions are true:

Predict I Scientist s knows the functional form of the outcome equation and its parameterization, and also knows that these are invariant over contexts (time and environment): $f^Y_{t'}(\cdot;\Theta_{t'}) = f^Y_t(\cdot;\Theta_t)$.

Predict II Scientist s knows the future direct effects of covariates, $\theta^0_{t'}$ $\theta^2_{t'}$, and $\theta^E_{t'} = 1$.

Predict III Scientist s knows how covariates will respond to treatment, $f^X_{t'}(D_{t'},\cdot;\Theta^X_{t'})$, $f^M_{t'}(D_{t'},\cdot;\Theta^M_{t'})$, and $f^E_{t'}(\cdot;\Theta^E_{t'})$.

Predict II allows scientist s to construct the predictions $\Pi_s(\theta^0_{t'}|t^*)$, $\Pi_s(\theta^2_{t'}|t^*)$, and $\Pi_s(\theta^E_{t'}|t^*) = 1$, and Predict III allows scientist s to construct the predictions $\Pi_s(X_{t'i(d)}|t^*)$ and $\Pi_s(E_{t'i(d)}|t^*)$. Now if scientist s possesses an estimate of the direct effect of treatment at past time t, $\widehat{\theta}^1_t$, Predict I instructs scientist s on how to combine these predictions with the direct effect of treatment in order to predict the total effect on the outcome variable at time t' in the future:

$$
\begin{aligned}
&\prod_s\big(\mathbb{E}_i[Y_{t'i(d')} - Y_{t'i(d)}]\,|\,t^*\big)\\
&\quad=\mathbb{E}_i\left[\prod_s(\theta^0_{t'}|t^*) + d'\,\widehat{\theta}^1_t + \prod_s(X_{t'i(d')}|t^*)\prod_s(\theta^2_{t'}|t^*) + \prod_s(E_{t'i(d')}|t^*)\right] \qquad \text{(Construction)}\\
&\qquad -\ \mathbb{E}_i\left[\prod_s(\theta^0_{t'}|t^*) + d\,\widehat{\theta}^1_t + \prod_s(X_{t'i(d)}|t^*)\prod_s(\theta^2_{t'}|t^*) + \prod_s(E_{t'i(d)}|t^*)\right]\\
&\quad=\mathbb{E}_i\left[\prod_s(\theta^0_{t'}|t^*) + d'\,\theta^1_{t'} + \prod_s(X_{t'i(d')}|t^*)\prod_s(\theta^2_{t'}|t^*) + \prod_s(E_{t'i(d')}|t^*)\right]\\
&\qquad -\ \mathbb{E}_i\left[\prod_s(\theta^0_{t'}|t^*) + d\,\theta^1_{t'} + \prod_s(X_{t'i(d)}|t^*)\prod_s(\theta^2_{t'}|t^*) + \prod_s(E_{t'i(d)}|t^*)\right] \qquad \text{(I+}\widehat{\theta}^1_t\text{)}\\
&\quad=\mathbb{E}_i\left[\prod_s(\theta^0_{t'}|t^*) + d'\,\theta^1_{t'} + X_{t'i(d')}\prod_s(\theta^2_{t'}|t^*) + E_{t'i(d')}\right]\\
&\qquad -\ \mathbb{E}_i\left[\prod_s(\theta^0_{t'}|t^*) + d\,\theta^1_{t'} + X_{t'i(d)}\prod_s(\theta^2_{t'}|t^*) + E_{t'i(d)}\right] \qquad \text{(III)}\\
&\quad= \mathbb{E}_i\left[\theta^0_{t'} + d'\,\theta^1_{t'} + \theta^2_{t'}X_{t'i(d')} + E_{t'i(d')}\right]\\
&\qquad\qquad - \mathbb{E}_i\left[\theta^0_{t'} + d\,\theta^1_{t'} + \theta^2_{t'}X_{t'i(d)} + E_{t'i(d)}\right] \qquad \text{(II)}\\
&\quad= \mathbb{E}_i[Y_{t'i(d')} - Y_{t'i(d)}], \qquad \text{(I)}
\end{aligned}
$$

where each equation is labeled by the assumption by which it is rendered an accurate prediction. The derivation of Π_s is analogous to the transport formula in Bareinboim and Pearl (2013a).

5.2 Prediction with Total Effects

Now suppose that scientist s possesses an estimate of the total effect of treatment $\mathbb{E}_i[\widehat{\triangle}_t^{D2}]$, obtained from data in the past at time t. In what circumstances will this causal effect be useful at the current time t^* for predicting how the outcome variable will change in response to a future intervention manipulating treatment from d to d'?

Total effects are useful under additional assumptions requiring that the behavior of covariates does not change between times t and t'. In addition to Predict I, II, and III, assume that at time t^*:

Predict IV Scientist s not only knows the direct effects of covariates in the future, but also knows that they will be the same as they were in the past, $\theta_{t'}^0 = \theta_t^0$, $\theta_{t'}^2 = \theta_t^1$, and $\theta_{t'}^E = \theta_t^E = 1$.

Predict V Scientist s not only knows how covariates will respond to treatment in the future, but also knows these equations will be the same as in the past, $f_{t'}^X(D_{t'}, \cdot; \Theta_{t'}^X) = f_t^X(D_t, \cdot; \Theta_t^X)$, $f_{t'}^M(D_{t'}, \cdot; \Theta_{t'}^M) = f_t^M(D_t, \cdot; \Theta_t^M)$, and $f_{t'}^E(\cdot; \Theta_{t'}^E) = f_t^E(\cdot; \Theta_t^E)$.

Thus the total effect identified in data from the past time t predicts the change from the future intervention:

$$\prod_s(\mathbb{E}_i[Y_{t'i(d')} - Y_{t'i(d)}]|t^*) = \mathbb{E}_i[\widehat{\triangle}_t^{D2}] \tag{Construction}$$

$$\equiv \mathbb{E}_i[Y_{ti(d')} - Y_{ti(d)}] \tag{Def D2}$$

$$= \mathbb{E}_i\Big\{\theta_t^0 + d'\,\theta_t^1 + X_{ti(d')}\theta_t^2 + E_{ti(d')} \;-\; \big[\theta_t^0 + d\,\theta_t^1 + X_{ti(d)}\theta_t^2 + E_{ti(d)}\big]\Big\} \tag{$\mathcal{D}_t$}$$

$$= \mathbb{E}_i\Big\{\theta_t^0 + d'\,\theta_t^1 + X_{t'i(d')}\theta_t^2 + E_{t'i(d')} \;-\; \big[\theta_t^0 + d\,\theta_t^1 + X_{t'i(d)}\theta_t^2 + E_{t'i(d)}\big]\Big\} \tag{V}$$

$$= \mathbb{E}_i\Big\{\theta_{t'}^0 + d'\,\theta_{t'}^1 + X_{t'i(d')}\theta_{t'}^2 + E_{t'i(d')} - \big[\theta_{t'}^0 + d\,\theta_{t'}^1 + X_{t'i(d)}\theta_{t'}^2 + E_{t'i(d)}\big]\Big\} \tag{IV}$$

$$= \mathbb{E}_i[Y_{t'i(d')} - Y_{t'i(d)}]. \tag{I}$$

Because they do not recover θ_t^1, total effects are used differently to construct the predictions than direct effects.

5.3 The Prediction Tradeoff Favoring Direct Effects

Predict IV and V cannot be invoked without invoking Predict II and III, which themselves are assumptions about knowledge of direct effects of covariates through the Θ's parameterizing the

selection equations. Predict I must also be invoked for prediction with total effects. Thus if denote by $\mathcal{P}t^*(\Delta_{t'})$ the subset of DGPs in $\{\mathcal{D}_t\}$ for which causal effects of type Δ_t can be used at the present time t^* to make accurate predictions about changes from interventions at future time t', we established above that

$$\mathcal{P}_{t^*}(\Delta_t^{D1}) = \left\{ \mathcal{D}_{t'} \;\middle|\; \text{Predict I, II, III} \right\}, \text{ while}$$

$$\mathcal{P}_{t^*}(\Delta_t^{D2}) = \left\{ \mathcal{D}_{t'} \;\middle|\; \text{Predict I, II, III, IV, V} \right\}.$$

Since Predict II and III are nested in Predict IV and V, Predict IV and V hold for a smaller subset of DGPs for which Predict II and III hold. This results in the prediction tradeoff:

$$\mathcal{P}_{t^*}(\Delta_{t'}^{D2}) \subset \mathcal{P}_{t^*}(\Delta_{t'}^{D1}).$$

Because it requires the stronger assumptions about the invariance of the behavior of covariates, prediction with total effects rules out circumstances of central interest in social systems that are ubiquitous. Returning to Figure 3 and our example DGP of rainfall and crop yields, we can see that any changes made to the behavior of covariates between times t and t' biases predictions made with total effects as constructed above. In Figure 3 one possibility is raised, which is that regulations pertaining to the use of pesticides changes between time t and t' due to human agency.

Electrons do not assemble and collectively decide to change their behavior. Nor do trees (Heckman (1997), Deaton (2010)). In contrast, one could easily imagine that between time t and t^*, say over the course of one year, a group of humans might have collectively decided to allow farmers to use certain productivity-enhancing pesticides on their crops. If farmers will in the future use pesticide to improve crop yield in years with little rain, but in the past did not do so because of regulation, now the context has changed between time t and t'.

Figure 3 also helps to illustrate why direct effects are so useful in social systems. A scientist who understands how the behavior of covariates has changed (in this case, through changes to f^M as relayed in Predict III) can still make an accurate prediction of outcomes after future interventions when they possess knowledge of direct effects. That is, as long as the scientist's prediction is accurate ($\Pi_s(E_{t'i(d)}|t^*) = E_{t'i(d)}$), they can move from line I+$\widehat{\theta}_t^1$ to line III in Section 5.1.

One could imagine many additional scenarios in which the behavior of covariates changes between times t and t'. The effectiveness of pesticides might improve between times t and t', violating Predict IV with $\theta_t^E \neq \theta_{t'}^E$ (again through human agency through research and development).[17] Predict IV might also be violated by the development of new hybrid crops that are more efficient, so that $\theta_t^1 \neq \theta_{t'}^1$. Alternatively, if water policy changes between times t and t', Predict V will be

[17] While plants and animals might change their behavior over time, the time span is so long relative to those of interest for questions addressed here that it is appropriate to assume these behaviors do not change. This stands in stark contrast to the behaviors of humans, especially in terms of collective decisions and/or their impacts on individual-level decisions.

violated as $f_t^X \neq f_{t'}^X$.

The point is that while assumptions similar to Predict IV and V may hold for physical systems, such assumptions are not likely to hold in social systems due to human agency. Choices of human beings make changes to the DGP ubiquitous in social systems, which therefore create any number of violations of assumptions about the temporal invariance of the process generating covariates. Total effects that require such strong restrictions on human behavior are unlikely to be useful for prediction in many social settings. Implications for the literature are discussed in the next section.

6 Implications for the Literature

6.1 An Example: Returns to Schooling

A tremendous amount of time, energy, and resources have been devoted to estimating the causal effects of educational attainment. To be explicit, the key reason for this focus is that policy makers and citizens might intervene to the DGP to encourage or discourage students from finishing high school and/or college. If we knew the changes from an intervention manipulating educational attainment D, it would help us to decide how much to spend as a society to implement that intervention.

Identifying causal effects of educational attainment is complicated by selection into treatment in response to an unobserved covariate (Card (2001), Belzil (2007)). Although overcoming selection into treatment is a non-trivial task (Angrist and Krueger (1991), Aliprantis (2012), Barua and Lang (2009)), suppose for the moment that social scientists had found empirical methods overcoming the unobserved nature of ability, and had identified the returns to schooling based on samples in which attainment were randomly assigned.

Would the total effects of education on earnings identified in past data be useful for predicting how wages would change in the future under interventions to increase educational attainment? Many mediators, or malleable covariates like E_t^M, could themselves respond to the assignment of treatment. Marginal Treatment Effects (MTEs) and Local Average Treatment Effects (LATEs) of educational attainment estimated in the literature like in Carneiro et al. (2011), Oreopoulos (2006a), and Oreopoulos (2006b) may reflect the direct effect of educational attainment, but due to selection into covariates may also reflect direct effects from any of the following causal variables:

Table 2: Covariates Whose Behavior Determines the Total Effect

	Evidence of Selection into Covariate in Response to Ed Attainment	Evidence of Effect of Covariate on Wages
Covariate	Study	Study
On-the-Job Training	Altonji and Blank (1999)	Brown (1989)
Job Training Program	–	Lee (2009)
Self-Employment	Blanchflower (2000)	Hamilton (2000)
Vocational Education	Bishop and Mane (2004)	Meer (2007)
Criminal Behavior	Jacob and Lefgren (2003)	Nagin and Waldfogel (1998)
Arrest	Grogger (1995)	Bushway (2004)
Incarceration	Lochner and Moretti (2004)	Kling (2006), Western et al. (2001)
Fertility	McCrary and Royer (2011)	Simonsen and Skipper (2006)
Household Formation	Nielsen and Svarer (2009)	Gemici (2011)
Geographic Location	Costa and Kahn (2000)	Baum-Snow and Pavan (2013), Black et al. (2009)
Military Service	Small and Rosenbaum (2008)	Angrist (1990)
Health (smoking)	Currie and Moretti (2003)	Auld (2005)
Working While in School	–	Light (2001)
Neighborhood Quality	–	Rosenbaum (1995), Aliprantis and Richter (2013)

This example helps to illustrate why total effects can be so weakly invariant in social settings. If the process generating the covariate were to change over time for any of these covariates, the total effects of attainment estimated in the literature would give biased predictions. Figure 4 displays DAGs of the total effects of educational attainment on wages at times t and t'. Suppose that at time t companies only provided on-the-job training to employees with certain levels of educational attainment. If this policy were to change between times t and t' so that at time t' companies provided training to all employees, regardless of their educational attainment level, then total effects would change.[18]

Accurate prediction with total effects requires there are not changes over time to the DGP related to *any* of the direct effects contributing to the total effect. It is difficult to imagine that the social processes related to *each* of the above mediators do not change in important ways over time.

[18]Similar examples can be found in Cartwright and Hardie (2012), who distinguish the evidence necessary to make accurate statements about the past from the evidence necessary to make accurate statements about the future. This can also be thought of as an example of how super exogeneity assumptions need not follow from weak exogeneity (Engle et al. (1983), Hendry and Richard (1982)).

A change in the direct effect of educational attainment on on-the-job training between times t and t' and the resulting realizations of $\mathcal{D}_t$ and $\mathcal{D}_{t'}$

M_1 = On-the-Job Training
M_2 = Job Training Program
M_3 = Self-Employment
M_4 = Vocational Education
M_5 = Criminal Behavior
M_6 = Arrest
M_7 = Incarceration
M_8 = Fertility
M_9 = Household Formation
M_{10} = Geographic Location
M_{11} = Military Service
M_{12} = Health (smoking)
M_{13} = Working While in School
M_{14} = Neighborhood Quality
Y = Wages

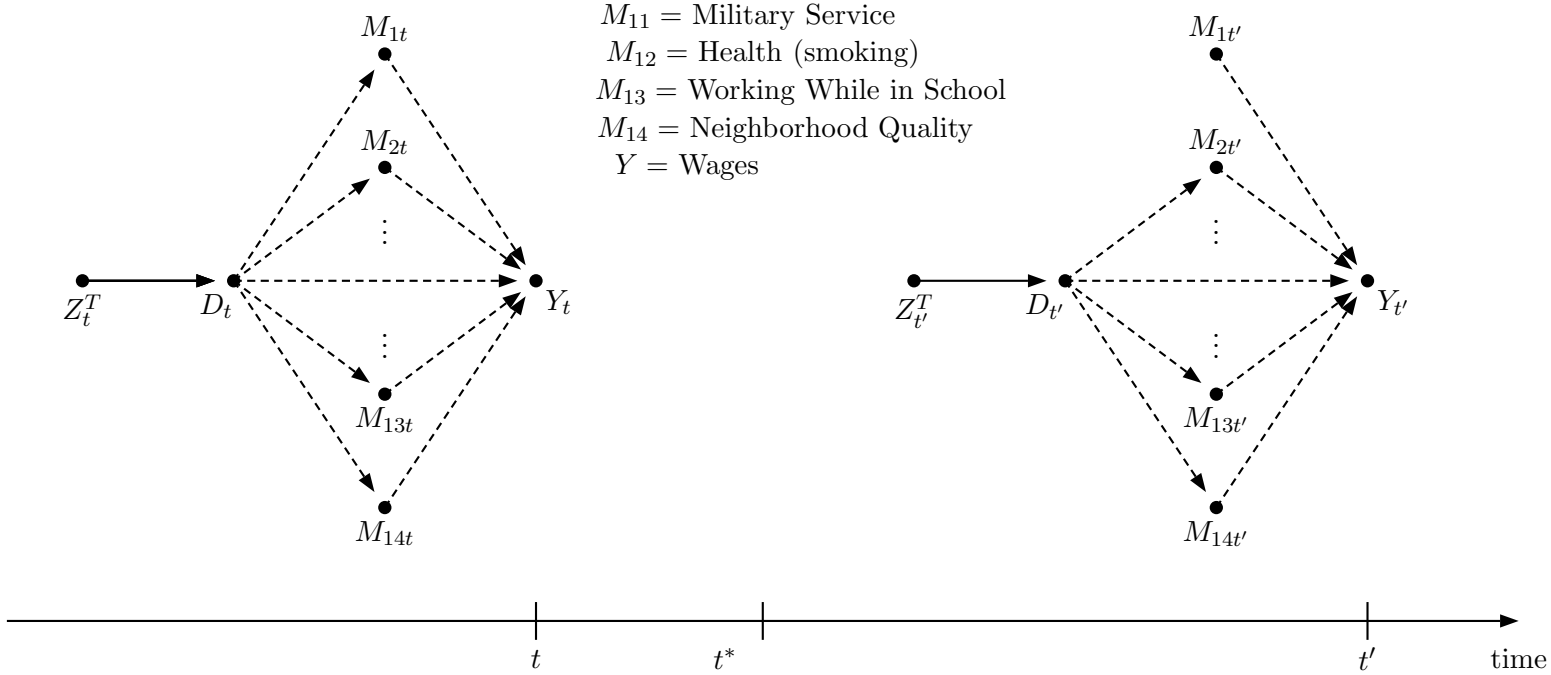


Figure 4: An example in which predictions of the future effect of $D_{t'}$ on $Y_{t'}$ (ie, of changes to $Y_{t'}$ from the intervention $Z_{t'}^T$) will be biased when constructed at current time t^* using total effects identified from past data (ie, from data collected at time t)

7 Conclusion

This paper studied a simple dynamic extension to the canonical static treatment effect framework. Treatment always influences the outcome variable in combination with other variables, which I refer to as covariates. I showed that there is a tradeoff between how easy it is to identify a causal effect in past data and its usefulness for predicting the future because covariates can respond even to a randomized treatment, and the behavior of covariates can change over time. I used the effects of precipitation on crop yields and of schooling on wages as examples to discuss why human agency is likely to change the behavior of covariates over time in many social systems.

References

Aliprantis, D. (2012). Redshirting, compulsory schooling laws, and educational attainment. *Journal of Educational and Behavioral Statistics 37*(2), 316–338.

Aliprantis, D. and F. G.-C. Richter (2013). Evidence of neighborhood effects from MTO: LATEs of neighborhood quality. *Mimeo., Federal Reserve Bank of Cleveland*.

Altonji, J. G. and R. M. Blank (1999). Race and gender in the labor market. In O. Ashenfeher and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3, pp. 3143–3259. North-Holland.

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from Social Security Administration records. *American Economic Review 80*(3), 313–335.

Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *The Economic Journal 114*(494), pp. C52–C83.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using Instrumental Variables. *Journal of the American Statistical Association 91*(434), 444–455.

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics 106*(4), 979–1014.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, J. D. and J.-S. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives 24*(2), 3–30.

Auld, M. C. (2005). Smoking, drinking, and income. *Journal of Human Resources 40*(2), 505–518.

Bareinboim, E. and J. Pearl (2013a). A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference 1*(1), 107–133.

Bareinboim, E. and J. Pearl (2013b, Forthcoming). Meta-transportability of causal effects: A formal approach. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Barua, R. and K. Lang (2009). School entry, educational attainment and quarter of birth: A cautionary tale of LATE. *NBER Working Paper 15236*.

Baum-Snow, N. and R. Pavan (2013, Forthcoming). Inequality and city size. *Review of Economics and Statistics*.

Belzil, C. (2007). The return to schooling in structural dynamic models: A survey. *European Economic Review 51*(5), 1059–1105.

Bishop, J. H. and F. Mane (2004). The impacts of career-technical education on high school labor market success. *Economics of Education Review 23*(4), 381–402.

Black, D., N. Kolesnikova, and L. Taylor (2009). Earnings functions when wages and prices vary by location. *Journal of Labor Economics 27*(1), 21–47.

Blanchflower, D. G. (2000). Self-employment in OECD countries. *Labour Economics 7*(5), 471 – 505.

Blundell, R. and M. C. Dias (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources 44*(3), 565–640.

Brown, J. N. (1989). Why do wages increase with tenure? on-the-job training and life-cycle wage growth observed within firms. *The American Economic Review*, 971–991.

Bushway, S. D. (2004). Labor market effects of permitting employer access to criminal history records. *Journal of Contemporary Criminal Justice 20*(3), 276–291.

Campbell, D. T. and J. Stanley (1963). *Experimental and Quasi-Experimental Designs for Research*. Cengage.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica 69*(5), 1127–1160.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011). Estimating marginal returns to education. *American Economic Review 101*(6), 2754–2781.

Carrell, S. E., B. I. Sacerdote, and J. E. West (2013). From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica 81*(3), 855–882.

Cartwright, N. and J. Hardie (2012). *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford University Press.

Chalak, K. and H. White (2011). An extended class of instrumental variables for the estimation of causal effects. *Canadian Journal of Economics 44*(1), 1–51.

Chalak, K. and H. White (2012). Causality, conditional independence, and graphical separation in settable systems. *Neural Computation* (24), 1611–1668.

Chen, B. and J. Pearl (2012). Regression and causation: A critical examination of econometrics textbooks. *Mimeo., UCLA Cognitive Systems Laboratory*.

Costa, D. L. and M. E. Kahn (2000). Power couples: Changes in the locational choice of the college educated, 1940-1990. *The Quarterly Journal of Economics 115*(4), 1287–1315.

Currie, J. and E. Moretti (2003). Mother's education and the intergenerational transmission of human capital: Evidence from college openings. *The Quarterly Journal of Economics 118*(4), 1495–1532.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424–455.

Deuchert, E. and M. Huber (2014). A cautionary tale about control variables in IV estimation with an application to health economics. *Mimeo., University of St. Gallen,*.

Duflo, E., R. Glennerster, and M. Kremer (2007). Using randomization in development economics research: A toolkit. In T. P. Schultz and J. A. Strauss (Eds.), *Handbook of Development Economics*, Volume 4, pp. 3895–3962. Elsevier.

Eberhardt, F. and R. Scheines (2007). Interventions and causal inference. *Philosophy of Science 74*(5), pp. 981–995.

Engle, R. F., D. F. Hendry, and J.-F. Richard (1983). Exogeneity. *Econometrica 51*(2), pp. 277–304.

Flores, C. A. and A. Flores-Lagunes (2010). Nonparametric partial identifcation of causal net and mechanism average treatment effects. *Mimeo., University of Miami*.

Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics 12*(2), 101–128.

Galles, D. and J. Pearl (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science 3*(1), 151–182.

Gemici, A. (2011). Family migration and labor market outcomes. *Mimeo., New York University*.

Goldberger, A. S. (1991). *A Course in Econometrics*. Harvard University Press.

Grogger, J. (1995). The effect of arrests on the employment and earnings of young men. *The Quarterly Journal of Economics 110*(1), 51–71.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica 12*, 1–118.

Hafeman, D. M. and T. J. VanderWeele (2011). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology 22*(6), 753–764.

Halpern, J. Y. (2000). Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research 12*(1), 317–337.

Hamilton, B. H. (2000). Does entrepreneurship pay? An empirical analysis of the returns to self-employment. *Journal of Political Economy 108*(3), 604–631.

Heckman, J. J. (1996). Randomization as an Instrumental Variable. *The Review of Economics and Statistics 78*(2), pp. 336–341.

Heckman, J. J. (1997). Instrumental Variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources 32*(3), 441–462.

Heckman, J. J. (2005). The scientific model of causality. *Sociological methodology 35*(1), 1–97.

Heckman, J. J. (2008). Econometric causality. *International Statistical Review 76*(1), 1–27.

Heckman, J. J. and S. Navarro (2007). Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics 136*(2), 341–396.

Heckman, J. J. and S. Navarro-Lozano (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics 86*(1), 30–57.

Heckman, J. J. and R. Pinto (2013). Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. *NBER Working Paper 19314*.

Heckman, J. J. and R. Pinto (2014). Causal analysis after Haavelmo. *Econometric Theory*. Forthcoming.

Heckman, J. J. and J. A. Smith (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives 9*(2), 85–110.

Heckman, J. J., S. Urzúa, and E. Vytlacil (2006). Understanding Instrumental Variables in models with essential heterogeneity. *The Review of Economics and Statistics 88*(3), 389–432.

Heckman, J. J. and E. Vytlacil (2001). Policy-relevant treatment effects. *The American Economic Review 91*(2), 107–111.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. Vytlacil (2007). Econometric evaluation of social programs, Part I: Causal models, structural models and econometric policy evaluation. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, Chapter 70, pp. 4779 – 4874. Elsevier.

Hendry, D. F. and J.-F. Richard (1982). On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics 20*(1), 3 – 33.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology 18*(1), 449–484.

Huber, M. (2013). Identifying causal mechanisms (primarily) based on Inverse Probability Weighting. *Journal of Applied Econometrics*. doi: 10.1002/jae.2341.

Imai, K., L. Keele, and D. Tingley (2010). A general approach to causal mediation analysis. *Psychological Methods 15*(4), 309–334.

Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of the Economic Literature 48*(2), 399–423.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of Local Average Treatment Effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 586.

Jacob, B. A. and L. Lefgren (2003). Are idle hands the devil's workshop? Incapacitation, concentration, and juvenile crime. *The American Economic Review 93*(5), pp. 1560–1577.

Jeffreys, H. (2011). *Scientific Inference*. Cambridge University Press.

Keane, M. P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics 156*(1), 3–20.

Keane, M. P. and K. I. Wolpin (2007). Exploring the usefulness of a nonrandom holdout sample for model validation: Welfare effects on female behavior. *International Economic Review 48*(4), 1351–1378.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *The American Economic Review 96*(3), 863–876.

Kydland, F. E. and E. C. Prescott (1977). Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy 85*(3), pp. 473–492.

Leamer, E. E. (2010). Tantalus on the road to Asymptopia. *The Journal of Economic Perspectives 24*(2), 31–46.

Lechner, M. and R. Miquel (2010). Identification of the effects of dynamic treatments by sequential conditional independence assumptions. *Empirical Economics 39*(1), 111–137.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies 76*(3), pp. 1071–1102.

Light, A. (2001). In-school work experience and the returns to schooling. *Journal of Labor Economics 19*(1), 65–93.

Lochner, L. and E. Moretti (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *The American Economic Review 94*(1), 155–189.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In K. Brunner and A. Meltzer (Eds.), *The Phillips Curve and Labor Markets*, Volume 1, pp. 19–46. Carnegie-Rochester Conference Series on Public Policy.

Manski, C. F. (2007). *Identification for Prediction and Decision*. Harvard University Press.

Mattei, A. and F. Mealli (2011). Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(5), 729–752.

McCrary, J. and H. Royer (2011). The effect of female education on fertility and infant health: Evidence from school entry laws using exact date of birth. *American Economic Review 101*(1), 158195.

Meer, J. (2007). Evidence on the returns to secondary vocational education. *Economics of education review 26*(5), 559–573.

Nagin, D. and J. Waldfogel (1998). The effect of conviction on income through the life cycle. *International Review of Law and Economics 18*(1), 25 – 40.

Nielsen, H. S. and M. Svarer (2009). Educational homogamy: How much is opportunities? *Journal of Human Resources 44*(4), 1066–1086.

Oreopoulos, P. (2006a). The compelling effects of compulsory schooling: Evidence from Canada. *The Canadian Journal of Economics / Revue canadienne d'Economique 39*(1), pp. 22–52.

Oreopoulos, P. (2006b). Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *The American Economic Review*, 152–175.

Pearl, J. (1993). On the statistical interpretation of structural equations. *Mimeo., UCLA Cognitive Systems Laboratory*.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420.

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys 3*, 96–146.

Pearl, J. (2009b). *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.

Pearl, J. (2013, Forthcoming). Trygve Haavelmo and the emergence of causal calculus. *Econometric Theory*.

Pearl, J. (2014, Forthcoming). Interpretation and identification of causal mediation. *Psychological Methods*.

Pearl, J. and E. Bareinboim (2011). Transportability of causal and statistical relations: A formal approach. *Proceedings of the Twenty-Fifth National Conference on Artificial Intelligence*, 247–254.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling 7*(9), 1393–1512.

Robins, J., T. Richardson, and P. Spirtes (2009). On identification and inference for direct effects. *Mimeo., Carnegie-Mellon University*.

Rogosa, D. (1987). Casual models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational and Behavioral Statistics 12*(2), 185–195.

Rosenbaum, J. E. (1995). Changing the geography of opportunity by expanding residential choice: Lessons from the Gautreaux program. *Housing Policy Debate 6*(1), 231–269.

Rosenbaum, P. R. (1984). The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General) 147*(5), pp. 656–666.

Rosenzweig, M. R. and K. I. Wolpin (2000). Natural "Natural Experiments" in Economics. *Journal of Economic Literature 38*, 827–874.

Schorfheide, F. and K. I. Wolpin (2013). To hold out or not to hold out. *Mimeo., University of Pennsylvania*.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Cengage.

Simonsen, M. and L. Skipper (2006). The costs of motherhood: an analysis using matching estimators. *Journal of Applied Econometrics 21*(7), 919–934.

Small, D. S. and P. R. Rosenbaum (2008). War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association 103*(483), 924–933.

Spirtes, P., R. Scheines, C. Glymour, T. Richardson, and C. Meek (2004). Causal inference. *Handbook of Quantitative Methodology in the Social Sciences*, 447–478.

Strotz, R. H. and H. O. Wold (1960). Recursive vs. nonrecursive systems: An attempt at synthesis. *Econometrica 28*(2), 417–427.

Todd, P. E. and K. I. Wolpin (2006). Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *The American Economic Review 96*(5), 1384–1417.

VanderWeele, T. J. (2009). Mediation and mechanism. *European Journal of Epidemiology 24*(5), 217–224.

Western, B., J. R. Kling, and D. F. Weiman (2001). The labor market consequences of incarceration. *Crime & delinquency 47*(3), 410–427.

White, H. and K. Chalak (2009). Settable systems: An extension of Pearl's Causal Model with optimization, equilibrium, and learning. *Journal of Machine Learning Research* (10), 1759–1799.

White, H. and K. Chalak (2013). Identification and identification failure for treatment effects using structural systems. *Econometric Reviews 32*(3), 273–317.

Wolpin, K. I. (2007). Ex ante policy evaluation, structural estimation, and model selection. *The American economic review 97*(2), 48–52.

Woodward, J. (2000). Explanation and invariance in the special sciences. *The British Journal for the Philosophy of Science 51*, 197–254.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Wooldridge, J. (2005). Violating ignorability of treatment by controlling for too many factors. *Econometric Theory 21*(05), 1026–1028.

Zellner, A. (2007). Philosophy and objectives of econometrics. *Journal of Econometrics 136*, 331–339.

8 Appendix A: Derivation of OLS and 2SLS Estimators

8.1 Notation for Matrix Algebra

For the sake of exposition, assume for these derivations that the constant term in the structural outcome equation $\theta_0 = 0$, and that the regressions are specified without constants. Additionally for the sake of exposition, recall that at the given level of measurement $\epsilon_{ti} \sim$ iid f^{ϵ}_{ti}, where f^{ϵ}_{ti} is simply the distribution of a random variable with finite variance, and no observable variables enter as argument of f^{ϵ}_{ti} at the given level of measurement. Since ϵ_{ti} is mean zero, has finite variance, and is independent of all observable variables, we can ignore it when taking expectations and constructing estimators. The ensuing analysis therefore considers DGPs omitting this variable.

Remember that $\boldsymbol{D}_t$ represents the $N \times 1$ vector of observations of D_{ti}. We also have N observations of $\boldsymbol{X}$ at both the time of assignment and the time of measurement, which were labeled as $\boldsymbol{X}_{t_0}$ and $\boldsymbol{X}_t$. Define the following $N \times 2$ vectors

$$\boldsymbol{W}_{t_0} \equiv [\boldsymbol{D}_t, \boldsymbol{X}_{t_0}], \qquad \boldsymbol{J}_{t_0} \equiv [\boldsymbol{Z}_t, \boldsymbol{X}_{t_0}]$$
$$\boldsymbol{W}_t \equiv [\boldsymbol{D}_t, \boldsymbol{X}_t], \text{ and } \qquad \boldsymbol{J}_t \equiv [\boldsymbol{Z}_t, \boldsymbol{X}_t].$$

Defining the $N \times 1$ and 2×1 vectors

$$\boldsymbol{Y}_t \equiv \begin{bmatrix} Y_{t1} \\ \vdots \\ Y_{tN} \end{bmatrix}, \qquad \boldsymbol{E}_t \equiv \begin{bmatrix} E_{t1} \\ \vdots \\ E_{tN} \end{bmatrix}, \qquad \boldsymbol{\theta}_t \equiv \begin{bmatrix} \theta^1_t \\ \theta^2_t \end{bmatrix},$$

it is possible to write the structural potential outcome Equation 7 in Section 3,

$$Y_{ti} \leftleftarrows D_{ti}\theta^1_t + X_{ti}\theta^2_t + E_{ti}, \tag{1}$$

as

$$\boldsymbol{Y}_t \leftleftarrows \boldsymbol{W}_t\boldsymbol{\theta}_t + \boldsymbol{E}_t.$$

Recall the regression Equations 8-10:

$$\boldsymbol{Y}_t = \boldsymbol{D}_t\alpha^1_t + \boldsymbol{H}_t \tag{5}$$
$$\boldsymbol{Y}_t = \boldsymbol{W}_t\boldsymbol{\beta}_t + \boldsymbol{K}_t \tag{6}$$
$$\boldsymbol{Y}_t = \boldsymbol{W}_{t_0}\boldsymbol{\gamma}_t + \boldsymbol{L}_t \tag{7}$$

8.2 Derivation of OLS Estimators

A little matrix algebra shows that:

$$\begin{aligned}\widehat{\alpha}_t^{1,OLS} &= (\boldsymbol{D}_t'\boldsymbol{D}_t)^{-1}[\boldsymbol{D}_t'\boldsymbol{Y}_t] \\ &= (\boldsymbol{D}_t'\boldsymbol{D}_t)^{-1}[\boldsymbol{D}_t'(\boldsymbol{D}_t\theta_t^1 + \boldsymbol{X}_t\theta_t^2 + \boldsymbol{E}_t)] \\ &= \theta_t^1 + \frac{\boldsymbol{D}_t'\boldsymbol{X}_t}{\boldsymbol{D}_t'\boldsymbol{D}_t}\theta_t^2 + \frac{\boldsymbol{D}_t'\boldsymbol{E}_t}{\boldsymbol{D}_t'\boldsymbol{D}_t}.\end{aligned}$$

Rewriting a ratio of the dot products of two $N \times 1$ vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ as

$$\frac{\boldsymbol{A}'\boldsymbol{B}}{\boldsymbol{B}'\boldsymbol{B}} = \frac{\frac{1}{N}\sum_{i=1}^N A_iB_i}{\frac{1}{N}\sum_{i=1}^N B_iB_i} \tag{21}$$

the Weak Law of Large Numbers implies that as N goes to infinity,

$$\widehat{\alpha}_t^{1,OLS} \xrightarrow{p} \theta_t^1 + \frac{\mathbb{E}\left[D_tX_t\right]}{\mathbb{E}\left[D_tD_t\right]}\theta_t^2 + \frac{\mathbb{E}\left[D_tE_t\right]}{\mathbb{E}\left[D_tD_t\right]}$$

as long as the above means are all finite.

Similarly,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_t^{OLS} &= (\boldsymbol{W}_t'\boldsymbol{W}_t)^{-1}[\boldsymbol{W}_t'\boldsymbol{Y}_t] \\ &= (\boldsymbol{W}_t'\boldsymbol{W}_t)^{-1}[\boldsymbol{W}_t'(\boldsymbol{W}_t\boldsymbol{\theta}_t + \boldsymbol{E}_t)] \\ &= \boldsymbol{\theta}_t + (\boldsymbol{W}_t'\boldsymbol{W}_t)^{-1}[\boldsymbol{W}_t'\boldsymbol{E}_t] \\ &= \boldsymbol{\theta}_t + \begin{bmatrix} \boldsymbol{D}_t'\boldsymbol{D}_t & \boldsymbol{D}_t'\boldsymbol{X}_t \\ \boldsymbol{X}_t'\boldsymbol{D}_t & \boldsymbol{X}_t'\boldsymbol{X}_t \end{bmatrix}^{-1} \left(\begin{bmatrix} D_{t1} & \cdots & D_{tN} \\ X_{t1} & \cdots & X_{tN} \end{bmatrix} \begin{bmatrix} E_{t1} \\ \vdots \\ E_{tN} \end{bmatrix} \right) \\ &= \boldsymbol{\theta}_t + \frac{1}{(\boldsymbol{D}_t'\boldsymbol{D}_t)(\boldsymbol{X}_t'\boldsymbol{X}_t) - (\boldsymbol{D}_t'\boldsymbol{X}_t)(\boldsymbol{X}_t'\boldsymbol{D}_t)} \begin{bmatrix} \boldsymbol{X}_t'\boldsymbol{X}_t & -\boldsymbol{D}_t'\boldsymbol{X}_t \\ -\boldsymbol{X}_t'\boldsymbol{D}_t & \boldsymbol{D}_t'\boldsymbol{D}_t \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_t'\boldsymbol{E}_t \\ \boldsymbol{X}_t'\boldsymbol{E}_t \end{bmatrix} \\ &= \begin{bmatrix} \theta_t^1 \\ \\ \theta_t^2 \end{bmatrix} + \begin{bmatrix} \frac{(\boldsymbol{X}_t'\boldsymbol{X}_t)(\boldsymbol{D}_t'\boldsymbol{E}_t)-(\boldsymbol{D}_t'\boldsymbol{X}_t)(\boldsymbol{X}_t'\boldsymbol{E}_t)}{(\boldsymbol{D}_t'\boldsymbol{D}_t)(\boldsymbol{X}_t'\boldsymbol{X}_t)-(\boldsymbol{D}_t'\boldsymbol{X}_t)(\boldsymbol{X}_t'\boldsymbol{D}_t)} \\ \\ \frac{-(\boldsymbol{X}_t'\boldsymbol{D}_t)(\boldsymbol{D}_t'\boldsymbol{E}_t)+(\boldsymbol{D}_t'\boldsymbol{D}_t)(\boldsymbol{X}_t'\boldsymbol{E}_t)}{(\boldsymbol{D}_t'\boldsymbol{D}_t)(\boldsymbol{X}_t'\boldsymbol{X}_t)-(\boldsymbol{D}_t'\boldsymbol{X}_t)(\boldsymbol{X}_t'\boldsymbol{D}_t)} \end{bmatrix}.\end{aligned}$$

Recalling Equation 21, as N goes to infinity

$$\widehat{\beta}_t^{1,OLS} \xrightarrow{p} \theta_t^1 + \frac{\mathbb{E}[X_tX_t]\,\mathbb{E}[D_tE_t] \;-\; \mathbb{E}[D_tX_t]\,\mathbb{E}[X_tE_t]}{\mathbb{E}[D_tD_t]\,\mathbb{E}[X_tX_t] \;-\; \mathbb{E}[D_tX_t]\,\mathbb{E}[X_tD_t]}$$

if the above means are finite.

And finally,

$$
\begin{aligned}
\widehat{\gamma}_t^{OLS} &= (\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})^{-1}[\boldsymbol{W}'_{t_0}\boldsymbol{Y}_t] \\
&= (\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})^{-1}[\boldsymbol{W}'_{t_0}(\boldsymbol{W}_t\boldsymbol{\theta}_t + \boldsymbol{E}_t)] \\
&= \frac{1}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0}) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} \begin{bmatrix} \boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0} & -\boldsymbol{D}'_t\boldsymbol{X}_{t_0} \\ -\boldsymbol{X}'_{t_0}\boldsymbol{D}_t & \boldsymbol{D}'_t\boldsymbol{D}_t \end{bmatrix} \begin{bmatrix} \boldsymbol{D}'_t\boldsymbol{D}_t\theta_t^1 + \boldsymbol{D}'_t\boldsymbol{X}_t\theta_t^2 + \boldsymbol{D}'_t\boldsymbol{E}_t \\ \boldsymbol{X}'_{t_0}\boldsymbol{D}_t\theta_t^1 + \boldsymbol{X}'_{t_0}\boldsymbol{X}_t\theta_t^2 + \boldsymbol{X}'_{t_0}\boldsymbol{E}_t \end{bmatrix}
\end{aligned}
$$

$$
= \begin{bmatrix} \frac{(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D}'_t\boldsymbol{D}_t)\theta_t^1 + (\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D}'_t\boldsymbol{X}_t)\theta_t^2 + (\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D}'_t\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0}) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} - \frac{(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)\theta_t^1 + (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{X}_t)\theta_t^2 + (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0}) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} \\ \frac{-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{D}_t)\theta_t^1 - (\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{X}_t)\theta_t^2 - (\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0}) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} + \frac{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)\theta_t^1 + (\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_t)\theta_t^2 + (\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0}) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} \end{bmatrix}
$$

$$
= \begin{bmatrix} \theta_t^1 + \frac{(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D}'_t\boldsymbol{X}_t) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{X}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0}) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)}\theta_t^2 + \frac{(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D}'_t\boldsymbol{E}_t) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0}) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} \\ \frac{-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{D}_t) + (\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0}) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)}\theta_t^1 + \frac{-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{X}_t) + (\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0}) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_A\boldsymbol{D}_t)}\theta_t^2 + \frac{-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{E}_t) + (\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0}) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} \end{bmatrix},
$$

so

$$
\begin{aligned}
\widehat{\gamma}_t^{1,OLS} \quad \xrightarrow{p} \quad \theta_t^1 &+ \frac{\mathbb{E}[X_{t_0}X_{t_0}]\,\mathbb{E}[D_tX_t] \;-\; \mathbb{E}[D_tX_{t_0}]\,\mathbb{E}[X_{t_0}X_t]}{\mathbb{E}[D_tD_t]\,\mathbb{E}[X_{t_0}X_{t_0}] \;-\; \mathbb{E}[D_tX_{t_0}]\,\mathbb{E}[X_{t_0}D_t]}\theta_t^2 \\
&+ \frac{\mathbb{E}[X_{t_0}X_{t_0}]\,\mathbb{E}[D_tE_t] \;-\; \mathbb{E}[D_tX_{t_0}]\,\mathbb{E}[X_{t_0}E_t]}{\mathbb{E}[D_tD_t]\,\mathbb{E}[X_{t_0}X_{t_0}] \;-\; \mathbb{E}[D_tX_{t_0}]\,\mathbb{E}[X_{t_0}D_t]}.
\end{aligned}
$$

8.3 Derivation of 2SLS Estimators

Similarly, we can perform some matrix algebra to see that

$$
\begin{aligned}
\widehat{\alpha}_t^{1,2SLS} &= (\boldsymbol{Z}'_t\boldsymbol{D}_t)^{-1}[\boldsymbol{Z}'_t\boldsymbol{Y}_t] \\
&= (\boldsymbol{Z}'_t\boldsymbol{D}_t)^{-1}[\boldsymbol{Z}'_t(\boldsymbol{D}_t\theta_t^1 + \boldsymbol{X}_t\theta_t^2 + \boldsymbol{E}_t)]
\end{aligned}
$$

$$
= \theta_t^1 + \frac{\boldsymbol{Z}'_t\boldsymbol{X}_t}{\boldsymbol{Z}'_t\boldsymbol{D}_t}\theta_t^2 + \frac{\boldsymbol{Z}'_t\boldsymbol{E}_t}{\boldsymbol{Z}'_t\boldsymbol{D}_t}, \tag{22}
$$

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_t^{2SLS} &= \Big[(\boldsymbol{W}_t'\boldsymbol{J}_t)(\boldsymbol{J}_t'\boldsymbol{J}_t)^{-1}(\boldsymbol{J}_t'\boldsymbol{W}_t)\Big]^{-1}\Big[(\boldsymbol{W}_t'\boldsymbol{J}_t)(\boldsymbol{J}_t'\boldsymbol{J}_t)^{-1}(\boldsymbol{J}_t'\boldsymbol{Y}_t)\Big] \\
&= \Big[(\boldsymbol{W}_t'\boldsymbol{J}_t)(\boldsymbol{J}_t'\boldsymbol{J}_t)^{-1}(\boldsymbol{J}_t'\boldsymbol{W}_t)\Big]^{-1}\Big[(\boldsymbol{W}_t'\boldsymbol{J}_t)(\boldsymbol{J}_t'\boldsymbol{J}_t)^{-1}(\boldsymbol{J}_t'\boldsymbol{W}_t\boldsymbol{\theta}_t)\Big] \\
&\quad + \Big[(\boldsymbol{W}_t'\boldsymbol{J}_t)(\boldsymbol{J}_t'\boldsymbol{J}_t)^{-1}(\boldsymbol{J}_t'\boldsymbol{W}_t)\Big]^{-1}\Big[(\boldsymbol{W}_t'\boldsymbol{J}_t)(\boldsymbol{J}_t'\boldsymbol{J}_t)^{-1}(\boldsymbol{J}_t'\boldsymbol{E}_t)\Big] \\
&= \boldsymbol{\theta}_t + \Big[(\boldsymbol{W}_t'\boldsymbol{J}_t)(\boldsymbol{J}_t'\boldsymbol{J}_t)^{-1}(\boldsymbol{J}_t'\boldsymbol{W}_t)\Big]^{-1}\Big[(\boldsymbol{W}_t'\boldsymbol{J}_M)(\boldsymbol{J}_t'\boldsymbol{J}_t)^{-1}(\boldsymbol{J}_t'\boldsymbol{E}_t)\Big],
\end{aligned} \tag{23}
$$

and

$$
\begin{aligned}
\widehat{\gamma}_t^{2SLS} &= \Big[(\boldsymbol{W}_{t_0}'\boldsymbol{J}_{t_0})(\boldsymbol{J}_{t_0}'\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}_{t_0}'\boldsymbol{W}_{t_0})\Big]^{-1}\Big[(\boldsymbol{W}_{t_0}'\boldsymbol{J}_{t_0})(\boldsymbol{J}_{t_0}'\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}_{t_0}'\boldsymbol{Y}_t)\Big] \\
&= \Big[(\boldsymbol{W}_{t_0}'\boldsymbol{J}_{t_0})(\boldsymbol{J}_{t_0}'\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}_{t_0}'\boldsymbol{W}_{t_0})\Big]^{-1}\Big[(\boldsymbol{W}_{t_0}'\boldsymbol{J}_{t_0})(\boldsymbol{J}_{t_0}'\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}_{t_0}'\boldsymbol{W}_t\boldsymbol{\theta}_t)\Big] \\
&\quad + \Big[(\boldsymbol{W}_{t_0}'\boldsymbol{J}_{t_0})(\boldsymbol{J}_{t_0}'\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}_{t_0}'\boldsymbol{W}_{t_0})\Big]^{-1}\Big[(\boldsymbol{W}_{t_0}'\boldsymbol{J}_{t_0})(\boldsymbol{J}_{t_0}'\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}_{t_0}'\boldsymbol{E}_M)\Big].
\end{aligned} \tag{24}
$$

Assuming for the sake of exposition that there is perfect compliance, so that $\boldsymbol{D}_t = \boldsymbol{Z}_t$, we can replace $\boldsymbol{J}_t = \boldsymbol{W}_t$. In this case, each of these 2SLS estimators reduces to their OLS counterpart, as:

$$
\begin{aligned}
\widehat{\alpha}_t^{1,2SLS} &= \theta_t^1 + \frac{\boldsymbol{Z}_t'\boldsymbol{X}_t}{\boldsymbol{Z}_t'\boldsymbol{D}_t}\theta_t^2 + \frac{\boldsymbol{Z}_t'\boldsymbol{E}_t}{\boldsymbol{Z}_t'\boldsymbol{D}_t} \\
&= \theta_t^1 + \frac{\boldsymbol{D}_t'\boldsymbol{X}_t}{\boldsymbol{D}_t'\boldsymbol{D}_t}\theta_t^2 + \frac{\boldsymbol{D}_t'\boldsymbol{E}_t}{\boldsymbol{D}_t'\boldsymbol{D}_t} \\
&= \widehat{\alpha}_t^{1,OLS},
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_t^{2SLS} &= \boldsymbol{\theta}_t + \Big[(\boldsymbol{W}_t'\boldsymbol{J}_t)(\boldsymbol{J}_t'\boldsymbol{J}_t)^{-1}(\boldsymbol{J}_t'\boldsymbol{W}_t)\Big]^{-1}\Big[(\boldsymbol{W}_t'\boldsymbol{J}_t)(\boldsymbol{J}_t'\boldsymbol{J}_t)^{-1}(\boldsymbol{J}_t'\boldsymbol{E}_t)\Big], \\
&= \boldsymbol{\theta}_t + (\boldsymbol{W}_t'\boldsymbol{W}_t)^{-1}(\boldsymbol{W}_t'\boldsymbol{E}_t) \\
&= \begin{bmatrix} \theta_t^1 \\ \\ \theta_t^2 \end{bmatrix} + \begin{bmatrix} \frac{(\boldsymbol{X}_t'\boldsymbol{X}_t)(\boldsymbol{D}_t'\boldsymbol{E}_t)-(\boldsymbol{D}_M'\boldsymbol{X}_t)(\boldsymbol{X}_t'\boldsymbol{E}_t)}{(\boldsymbol{D}_t'\boldsymbol{D}_t)(\boldsymbol{X}_t'\boldsymbol{X}_t)-(\boldsymbol{D}_t'\boldsymbol{X}_t)(\boldsymbol{X}_t'\boldsymbol{D}_t)} \\ \\ \frac{-(\boldsymbol{X}_M'\boldsymbol{D}_t)(\boldsymbol{D}_t'\boldsymbol{E}_t)+(\boldsymbol{D}_t'\boldsymbol{D}_t)(\boldsymbol{X}_t'\boldsymbol{E}_t)}{(\boldsymbol{D}_M'\boldsymbol{D}_t)(\boldsymbol{X}_t'\boldsymbol{X}_t)-(\boldsymbol{D}_t'\boldsymbol{X}_t)(\boldsymbol{X}_t'\boldsymbol{D}_t)} \end{bmatrix} \\
&= \widehat{\boldsymbol{\beta}}_t^{OLS},
\end{aligned}
$$

and

$$
\begin{aligned}
\widehat{\boldsymbol{\gamma}}_t^{2SLS} &= \left[(\boldsymbol{W}'_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J}'_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}'_{t_0}\boldsymbol{W}_{t_0})\right]^{-1}\left[(\boldsymbol{W}'_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J}'_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}'_{t_0}\boldsymbol{W}_M\boldsymbol{\theta}_t)\right] \\
&\quad + \left[(\boldsymbol{W}'_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J}'_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}'_{t_0}\boldsymbol{W}_{t_0})\right]^{-1}\left[(\boldsymbol{W}'_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J}'_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}'_{t_0}\boldsymbol{E}_t)\right] \\
&= \left[(\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})(\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})^{-1}(\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})\right]^{-1}\left[(\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})(\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})^{-1}(\boldsymbol{W}'_{t_0}\boldsymbol{W}_t\boldsymbol{\theta}_t)\right] \\
&\quad + \left[(\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})(\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})^{-1}(\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})\right]^{-1}\left[(\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})(\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})^{-1}(\boldsymbol{W}'_{t_0}\boldsymbol{E}_t)\right]
\end{aligned}
$$

$$
= \begin{bmatrix} \theta_t^1 + \frac{(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D}'_t\boldsymbol{X}_t)-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{X}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)}\theta_t^2 + \frac{(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D}'_t\boldsymbol{E}_t)-(\boldsymbol{D}'_M\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} \\ \frac{-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_M)(\boldsymbol{D}'_t\boldsymbol{D}_t)+(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)}\theta_t^1 + \frac{-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{X}_t)+(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)}\theta_t^2 + \frac{-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{E}_t)+(\boldsymbol{D}'_M\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_{ti})} \end{bmatrix}
$$

$$
= \widehat{\boldsymbol{\gamma}}_t^{OLS}.
$$

9 Appendix B: Monte Carlo Evidence on Identification

This Appendix uses Monte Carlo simulations to illustrate identification of causal effects in four specific DGPs in $\{\mathcal{D}_t\}$. DGP $\mathcal{D}_t^I$ is characterized by X_{ti} and E_{ti} not responding to the values drawn of D_{ti}. We might also think that agents make choices in response to the values of causal variables on which they have information. For example, the DGP might be characterized by farmers choosing to irrigate their crops in years with little precipitation ($\mathcal{D}_t^{II}$); choosing to apply more pesticide to their crops in years with little precipitation ($\mathcal{D}_t^{III}$); or choosing to irrigate their crops and apply more pesticide to them in years with little precipitation ($\mathcal{D}_t^{IV}$). It is important to note here that it is the farmer who is making active choices in DGPs $\mathcal{D}_t^{II}$-$\mathcal{D}_t^{IV}$, and not the crops or the plots of lands.

Table 3 presents Monte Carlo results showing the performance of the OLS estimators from Equations 1-7 when estimated on 100,000 simulated data points generated from these DGPs. The precise parameterizations are as follows: The potential outcome equation is the same across all simulated DGPs:

$$\begin{aligned} Y_{ti} &\leftleftarrows \theta_t^0 + D_{ti}\theta_t^1 + X_{ti}\theta_t^2 + E_{ti} \\ &\leftleftarrows 2.0 + D_{ti} \cdot 1.0 + X_{ti} \cdot 1.0 + E_{ti}. \end{aligned}$$

As well, in all simulated DGPs treatment is randomized with

$$\begin{aligned} D_{ti} \leftleftarrows 0.5Z_{ti}^T + 0.5U_{ti}^D \quad & \text{where} \\ & Z_{ti}^T \sim iidU[-1,1] \ \text{ and} \\ & U_{ti}^D \sim iidU[-1,1], \end{aligned} \tag{2*}$$

stated equivalently as D_{ti} being an *iid* random variable that follows the triangle distribution with lower limit -1, upper limit 1, and mode 0.

In $\mathcal{D}_t^I$ the remaining selection equations 3, 1, 4, and 5 are such that:

$$E_{ti}^P = 0 \ \ \forall \ \ i$$

with

$$X_{ti} \sim U[-\tfrac{1}{2}, \tfrac{1}{2}], \qquad \text{and} \qquad E_{ti} \equiv E_{ti}^P + E_{ti}^M \sim U[-\tfrac{1}{2}, \tfrac{1}{2}]. \tag{DGP I}$$

Let $U_{ti}^X \sim iidU[0,1]$. In $\mathcal{D}_t^{II}$ all features of the model are the same as in $\mathcal{D}_t^I$ except that observed covariates are selected in response to treatment:

$$X_{ti} \leftleftarrows \begin{cases} 1 - D_{ti} & \text{if } U_{ti}^X \leq 0.5 \\ A_{ti} & \text{otherwise, where } A_{ti} \sim U[-\tfrac{1}{2}, \tfrac{1}{2}] \end{cases} \tag{DGP II}$$

Similarly, $\mathcal{D}_t^{III}$ is the same as $\mathcal{D}_t^I$ except that now unobserved covariates are selected in response

to treatment. Unobserved covariates are a combination of permanent and malleable factors:

$$E_{ti} = 0.25E_{ti}^{P} + 0.75E_{ti}^{M}.$$

The permanent factor is determined as $E_{ti}^{P} \sim U[-\frac{1}{2}, \frac{1}{2}]$, and letting $U_{ti}^{M} \sim iidU[0,1]$, the malleable factor is determined in response to treatment as:

$$E_{ti}^{M} \leftleftarrows \begin{cases} 1 - D_{ti} & \text{if } U_{ti}^{M} \leq 0.75 \\ B_{ti} & \text{otherwise, where } B_{ti} \sim iidU[-\frac{1}{2}, \frac{1}{2}] \end{cases} \tag{DGP III}$$

Finally, $\mathcal{D}_t^{IV}$ is the same as $\mathcal{D}_t^{I}$ except that observed covariates are selected in response to treatment as in DGP $\mathcal{D}_t^{II}$ and unobserved covariates are selected in response to treatment as in $\mathcal{D}_t^{III}$.

9.1 Direct and Total Effects Represent Information about Distinct Counterfactuals from the Same DGP

Table 3 shows the remarkable robustness of the identification of the average total effect. The OLS estimators from Equations 8 and 10, $\widehat{\alpha}_t^{1,OLS}$ and $\widehat{\gamma}_t^{1,OLS}$, identify the average total effect in all four DGPs.

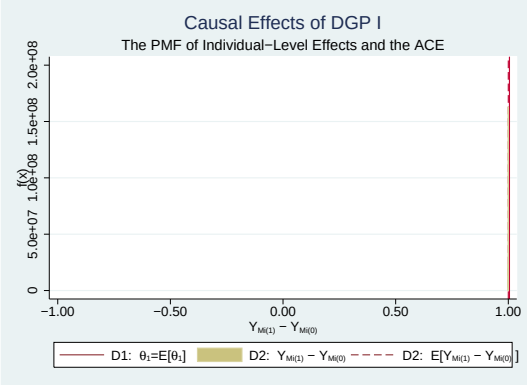
The causal effect parameters of these DGPs can be seen in Figure 5 and in the first two rows of Table 3. Both Figure 5 and Table 3 illustrate that direct and total causal effects represent information about distinct counterfactuals from the same DGP (Galles and Pearl (1998), Halpern (2000), Pearl (2013), Pearl (2009b)). Recalling definitions D1 and D2 of direct and total causal effects, Figure 5 helps to illustrate that direct effects represent information about the counterfactual quantity $\mathbb{E}[Y_t|do(D_t = d, X_t = x, E_t = e)]$, which holds covariates fixed. In contrast, total effects represent information about the counterfactual quantity $\mathbb{E}[Y_t|do(D_t = d)]$, which allows individuals to select into covariates freely. Alternatively, direct effects represent information about the counterfactual $Y_{ti(D=d,X=x,E=e)}$ and total effects represent information about the counterfactual $Y_{ti(D=d)}$.

Focusing on Figure 5, we can see that in DGP $\mathcal{D}_t^{I}$ all effect parameters coincide because there is no selection into covariates. In DGPs $\mathcal{D}_t^{II}$ and $\mathcal{D}_t^{III}$ there is heterogeneity in individual-level total effects, and the average net effect is the population-weighted average of these two discrete individual-level effects. In DGP $\mathcal{D}_t^{IV}$ there is even more heterogeneity in individual-level total effects, and the average total effect is actually negative. Since the direct effect θ_t^1 is constant across individuals in all DGPs, the only source of heterogeneity of individual-level effects in DGPs $\mathcal{D}_t^{II}$-$\mathcal{D}_t^{IV}$ is selection into covariates (Equations 26 and 27), not heterogeneous direct effects (Equation 25).

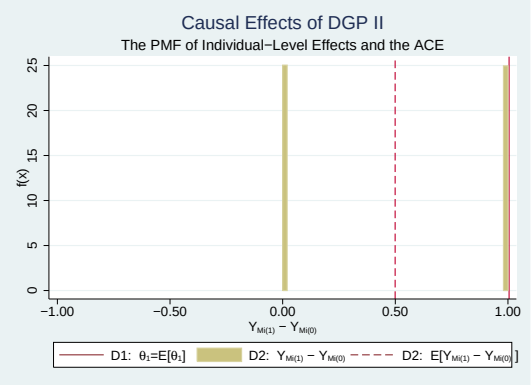
Table 3: Estimation Results on Data Simulated from Data Generating Processes with Various Selection Rules

		A DGP $\mathcal{D}_t$ is Fully Specified by: -The Potential Outcome Equation -Selection Equations f_t^D, f_t^X, f_t^P, f_t^M, and f_t^E			
		DGP $\mathcal{D}_t^I$	DGP $\mathcal{D}_t^{II}$	DGP $\mathcal{D}_t^{III}$	DGP $\mathcal{D}_t^{IV}$
	Potential Outcomes:	$Y_{ti} \leftleftarrows \theta_t^0 + \theta_t^1 D_{ti} + \theta_t^2 X_{ti} + E_{ti}$			
	Selection Rule f_t^X:	$X_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}]$	$X_{ti} \leftleftarrows f_t^X(D_{ti})$	$X_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}]$	$X_{ti} \leftleftarrows f_t^X(D_{ti})$
	Selection Rule f_t^E:	$E_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}]$	$E_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}]$	$E_{ti} \leftleftarrows f_t^E(D_{ti})$	$E_{ti} \leftleftarrows f_t^E(D_{ti})$
Randomized D_t, f_t^D: $D_{ti} \leftleftarrows 0.5Z_{ti}^T + 0.5U_{ti}^D$, $Z_{ti}^T, U_{ti}^D \sim iidU[-1,1]$	**Causal Effects**				
	D1: θ_t^1	1.00	1.00	1.00	1.00
	D2: $\mathbb{E}[Y_{ti(D_{ti}=1)} - Y_{ti(D_{ti}=0)}]$	1.00	0.50	0.44	–0.06
	Estimate				
	$\widehat{\alpha}_t^{1,OLS}$ $(\xrightarrow{p} \widehat{\alpha}_t^{1,2SLS})$	1.00	0.49	0.44	–0.06
	$\widehat{\beta}_t^{1,OLS}$ $(\xrightarrow{p} \widehat{\beta}_t^{1,2SLS})$	1.01	1.00	0.44	0.43
	$\widehat{\gamma}_t^{1,OLS}$ $(\xrightarrow{p} \widehat{\gamma}_t^{1,2SLS})$	1.00	0.49	0.44	–0.06
	Exclusion Restrictions				
	D1: $\mathbb{E}[Z_{ti}^T E_{ti}]$	0.00	0.00	–0.09	0.12
	D2: $CORR(Y_{ti}, Z_{ti}^T)\|D_{ti}, E_{ti}^P$	0.00	0.00	–0.01	0.01
	Selection into Covariates				
	$\mathbb{E}[X_{ti}\|D_{ti} > 0]$	0.00	0.33	0.00	0.28
	$\mathbb{E}[X_{ti}\|D_{ti} < 0]$	0.00	0.67	0.00	0.62
	$\mathbb{E}[E_{ti}\|D_{ti} > 0]$	0.00	0.00	0.38	0.31
	$\mathbb{E}[E_{ti}\|D_{ti} < 0]$	0.00	0.00	0.75	0.71

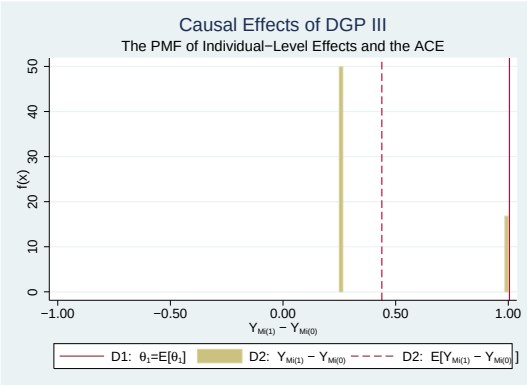
Note: The specified DGPs were used to generate 100,000 observations. The precise functions f_t^X and f_t^E used in each simulated DGP are specified in Section 4 in the text.



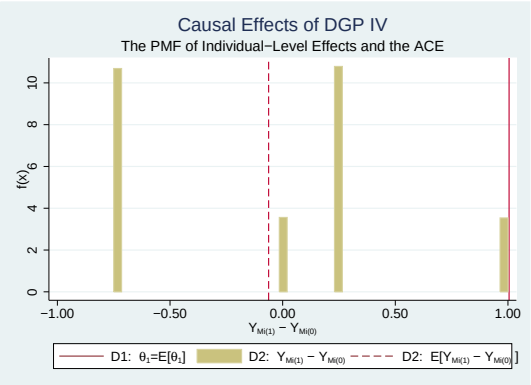
(a) DGP I



(b) DGP II

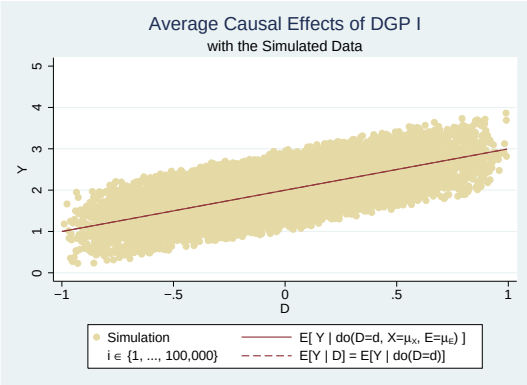


(c) DGP III

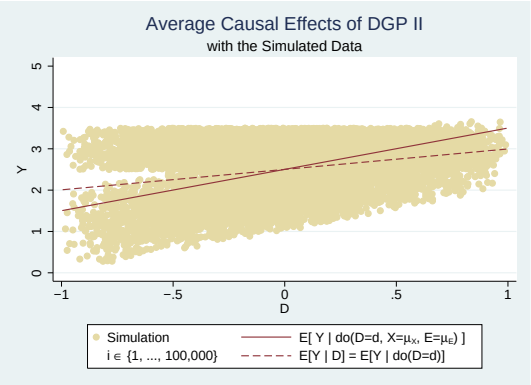


(d) DGP IV

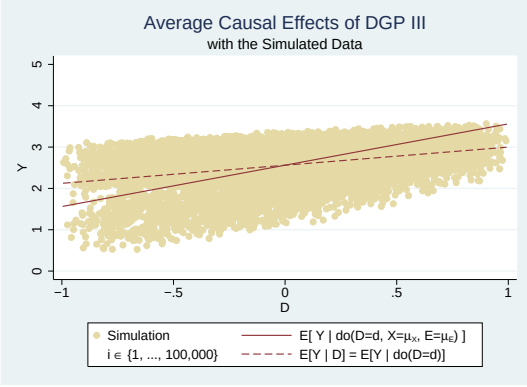
Figure 5: Causal Effects of Data Generating Processes $\mathcal{D}_t^I$-$\mathcal{D}_t^{IV}$



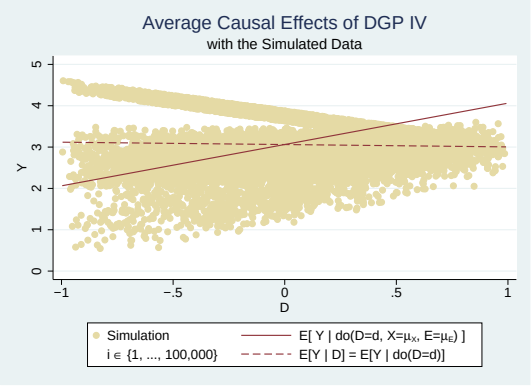
(a) DGP I



(b) DGP II



(c) DGP III



(d) DGP IV

Figure 6: Coefficients of Data Generating Processes $\mathcal{D}_t^I$-$\mathcal{D}_t^{IV}$

9.2 Debates about Alternative Approaches to Causal Inference: One Person's Bias is Another Person's Identification

DGPs $\mathcal{D}_t^{II}$-$\mathcal{D}_t^{IV}$ illustrate that one person's bias is another person's identification, due to the fact that the exclusion restriction identifying direct effects is distinct from the exclusion restriction identifying total effects.

In DGP $\mathcal{D}_t^{II}$ what represents bias for the researcher trying to identify the total effect represents identification of the direct effect. This DGP is helpful as a guide for researchers' decision of whether to control for observed covariates or not. When it is suspected that D_{ti} is an argument in selection Equation 3 as in DGP $\mathcal{D}_t^{II}$, Angrist and Pischke (2009) recommend estimating Equation 8 (p 66) and Duflo et al. (2007) (p 3949) and Rosenbaum (1984) recommend estimating Equation 10. Table 3 shows that both of these estimators will identify the average total effect of treatment, $\mathbb{E}_i[\triangle_{ti}^{D2}]$. Should the researcher control for the observed covariates determined in response to treatment and estimate Equation 9, $\widehat{\beta}_t^{1,OLS}$ will be a biased estimator of the average total effect $\mathbb{E}_i[\triangle_{ti}^{D2}]$ (Wooldridge (2005), Heckman and Navarro-Lozano (2004)). Chalak and White (2011) refer to this as "included variable bias." At the same time, though, $\widehat{\beta}_t^{1,OLS}$ will identify the direct causal effect θ_t^1 in $\triangle_{ti}^{D1}$.

In DGPs $\mathcal{D}_t^{III}$ and $\mathcal{D}_t^{IV}$ what represents bias for the researcher trying to identify the direct effect represents identification of the total effect. In light of DGPs $\mathcal{D}_t^{III}$ and $\mathcal{D}_t^{IV}$, previous criticisms of the experimentalist approach can be seen as discussions of identification using DGPs with random variation in treatment that impacts outcomes through covariates. One of Heckman (1997)'s concerns about the total effects identified in Angrist (1990) is that they cannot distinguish between the direct effect of treatment and the direct effect of unobserved covariates selected in response to the quasi-randomly assigned treatment. The concerns raised in Rosenzweig and Wolpin (2000) and Keane (2010) about total effects identified by the quasi-random assignment of treatment generated by natural experiments are likewise related to selection into covariates, creating a difference between the total effect and direct effect identified by IV estimators. Finally, the distinction between exogeneity and orthogonality made in Deaton (2010) can be seen as a distinction between orthogonality conditions made at two points in time, the time of assignment (t_0) and the time of measurement (t). Deaton's concern is that even if orthogonality conditions hold for a given DGP at time t_0, the later ones at t necessary for identification can be violated due to selection into covariates.[19] Further discussion of the cases when conditioning and setting/fixing variables coincide can be found in Heckman and Pinto (2014).

[19] A similar point about DGPs in which the direct effect is not identified is made in White and Chalak (2013). Further discussions on the limitations of effects identified by randomized treatments are provided in Leamer (2010) and Heckman and Smith (1995).

10 Appendix C: The Marginal Treatment Effect and Sources of Effect Heterogeneity

Here we also present a brief discussion of the Marginal Treatment Effect (MTE) in recognition of the fact that all varieties of the total effects in the program evaluation literature can be written as weighted averages of MTEs (Heckman and Vytlacil (2005)). Suppose that we allow for heterogeneous responses to treatment, replacing Equation 1* with

$$Y_{ti} \leftleftarrows \theta_t^0 + D_{ti}\theta_{ti}^1 + X_{ti}(D_{ti})\theta_t^2 + E_{ti}(D_{ti}). \tag{1*}$$

Define the Marginal Treatment Effect (MTE) as:

$$\triangle_t^{MTE}(d, d', U_{ti}^D) \;\equiv\; \mathbb{E}_i[\, Y_{ti(d')} - Y_{ti(d)} \;\mid\; U_{ti}^D \,].$$

Note that this definition of MTE contrasts with that in Heckman and Vytlacil (2005) and the generalizations in Heckman et al. (2006) in that it does not condition on observed covariates X_{ti}.

Under the specification of the DGP in Equation 1* one possible reason for heterogeneity in $\triangle_t^{MTE}$ over the support of U_{ti}^D is a correlation between the unmeasured factors driving selection into D_{ti} (U_{ti}^D in Equation 2) and the direct effect of D_{ti} (θ_{ti}^1). In this case the expected causal effect of D_{ti} (under all definitions) might vary due to idiosyncratic differences in the direct effect of treatment over different parts of the support of U_{ti}^D. Different interventions might induce individuals with different U_{ti}^D to select into D_{ti}, generating heterogeneous MTEs through heterogeneous response to treatment:

$$\mathbb{E}_i\Big[\theta_{ti}^1 \,(\; d' - d \;)\;\Big|\; U_{ti}^D\Big]. \tag{25}$$

Another possible reason for heterogeneity in $\triangle_t^{MTE}$ over the support of U_{ti}^D is that either observed or unobserved covariates, $X_{ti}(D_{ti})$ or $E_{ti}(D_{ti})$ respectively, are determined in response to treatment, so that:

$$\mathbb{E}_i\Big[\theta_{ti}^2 \,(\; X_{ti(d')} - X_{ti(d)} \;)\;\Big|\; U_{ti}^D\Big] \neq 0 \;\;,\text{ or} \tag{26}$$

$$\mathbb{E}_i\Big[(\; E_{ti(d')} - E_{ti(d)} \;)\;\Big|\; U_{ti}^D\Big] \neq 0 \;. \tag{27}$$

11 Appendix D: Extension to DGPs with Selection into Treatment in Response to an Unobserved Covariate

Table 4 presents Monte Carlo results showing the performance of the OLS estimators from Equations 8-10 as well as their 2SLS analogues when DGPs $\mathcal{D}_t^I$-$\mathcal{D}_t^{IV}$ also exhibit selection into treatment. These DGPs are characterized by the following structural equations:

$$E_{ti}^P \leftleftarrows f^P(U_{ti}^P;\ \Theta^P) \tag{28}$$
$$D_{ti} \leftleftarrows f^D(Z_{ti}^T, E_{ti}^P, U_{ti}^D\ ;\ \Theta^D) \tag{29}$$
$$X_{ti} \leftleftarrows f^X(D_{ti}, E_{ti}^P, U_{ti}^X\ ;\ \Theta^X) \tag{30}$$
$$E_{ti}^M \leftleftarrows f^M(D_{ti}, X_{ti}, U_{ti}^M;\ \Theta^M) \tag{31}$$
$$E_{ti} \leftleftarrows f^E(E_{ti}^P, E_{ti}^M\ ;\ \Theta^E) \tag{32}$$
$$Y_{ti} \leftleftarrows \theta_t^0 + D_{ti}\theta_t^1 + X_{ti}\theta_t^2 + E_{ti}. \tag{33}$$

In terms of specification, the potential outcome equation is still the same across all DGPs:

$$Y_{ti} \leftleftarrows 2.0 + D_{ti} \cdot 1.0 + X_{ti} \cdot 1.0 + E_{ti}.$$

The difference is that now, in all simulated DGPs treatment is selected according to

$$D_{ti} \leftleftarrows 0.5Z_{ti}^T + 0.25U_{ti}^D + 0.25E_{ti}^P$$

where both $Z_{ti}^T, U_{ti}^D \sim iidU[-1,1]$. E_{ti}^P represents a permanent component of the unobserved covariate and E_{ti}^M represents a malleable component of the unobserved covariate as follows:

$$E_{ti} \leftleftarrows 0.25E_{ti}^P + 0.75E_{ti}^M.$$

In DGP $\mathcal{D}_t^I$ both $E_{ti}^P, E_{ti}^M \sim iidU[-\frac{1}{2}, \frac{1}{2}]$ and the remaining selection equation is specified to be:

$$X_{ti} \sim U[-\tfrac{1}{2}, \tfrac{1}{2}].$$

In DGP $\mathcal{D}_t^{II}$ all features of the model are the same as in DGP $\mathcal{D}_t^{II}$ except that observed covariates are selected in response to treatment:

$$X_{ti} \leftleftarrows \begin{cases} 1 - D_{ti} & \text{if } U_{ti}^X \leq 0.5 \\ A_{ti} & \text{otherwise, where } A_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}] \end{cases}$$

where $U_{ti}^X \sim iidU[0,1]$.

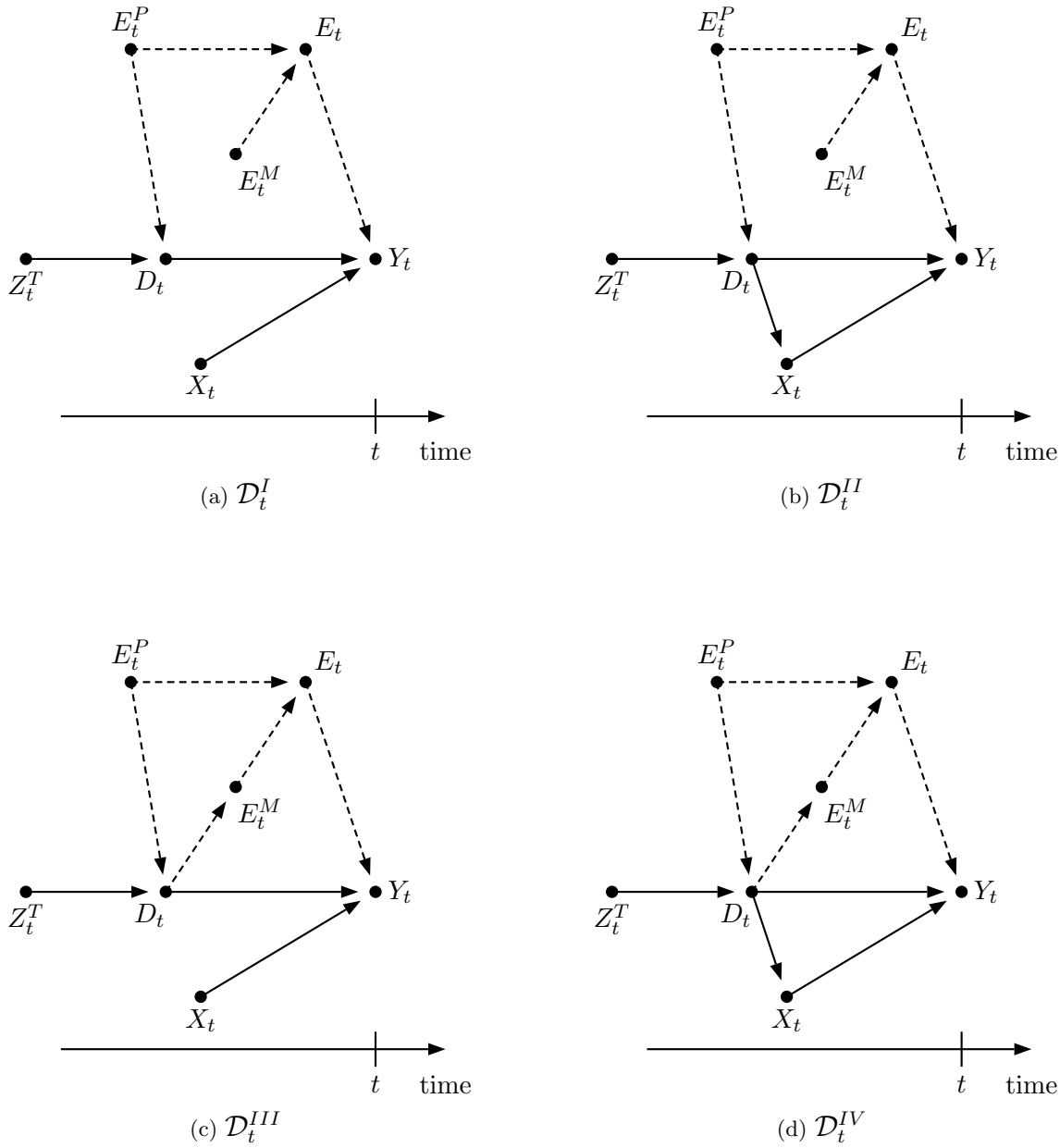
Similarly, DGP $\mathcal{D}_t^{III}$ is the same as DGP $\mathcal{D}_t^I$ except that now unobserved covariates are selected

in response to treatment as

$$E_{ti}^M \leftleftarrows \begin{cases} 1 - D_{ti} & \text{if } U_{ti}^E \leq 0.75; \\ B_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}] & \text{if } U_{ti}^E > 0.75, \end{cases}$$

with $U_{ti}^E \sim U[0, 1]$.

Finally, DGP $\mathcal{D}_t^{IV}$ is the same as DGP $\mathcal{D}_t^I$ except that observed covariates are selected in response to treatment as in DGP $\mathcal{D}_t^{II}$ and unobserved covariates are selected in response to treatment as in DGP $\mathcal{D}_t^{III}$.



(a) $\mathcal{D}_t^I$

(b) $\mathcal{D}_t^{II}$

(c) $\mathcal{D}_t^{III}$

(d) $\mathcal{D}_t^{IV}$

Figure 7: Directed Acyclic Graphs of Four Data Generating Processes from $\{\mathcal{D}_t\}$ Including the Total Effect Intervention Z_t^T

Table 4: Estimation Results on Data Simulated from Data Generating Processes with Selection into Treatment

	A DGP $\mathcal{D}_t$ is Fully Specified by: -The Potential Outcome Equation -Selection Equations f^D, f^X, and f^E			
	DGP $\mathcal{D}_t^I$	DGP $\mathcal{D}_t^{II}$	DGP $\mathcal{D}_t^{III}$	DGP $\mathcal{D}_t^{IV}$
Potential Outcomes:	$Y_{ti} \leftleftarrows \theta_t^0 + \theta_t^1 D_{ti} + \theta_t^2 X_{ti} + E_{ti}$			
Selection Rule f_t^X:	$X_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}]$	$X_{ti} \leftleftarrows f_t^X(D_{ti})$	$X_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}]$	$X_{ti} \leftleftarrows f_t^X(D_{ti})$
Selection Rule f_t^E:	$E_{ti}^M \sim U[-\frac{1}{2}, \frac{1}{2}]$	$E_{ti}^M \sim U[-\frac{1}{2}, \frac{1}{2}]$	$E_{ti}^M \leftleftarrows f_t^E(D_{ti})$	$E_{ti}^M \leftleftarrows f_t^E(D_{ti})$
Causal Effects				
D1: θ_t^1	1.00	1.00	1.00	1.00
D2: $\mathbb{E}[Y_{ti(D_{ti}=1)} - Y_{ti(D_{ti}=0)}]$	1.00	0.50	0.44	–0.06
Estimate				
$\widehat{\alpha}_t^{1,OLS}$	1.04	0.54	0.49	–0.02
$\widehat{\beta}_t^{1,OLS}$	1.05	1.05	0.49	0.49
$\widehat{\gamma}_t^{1,OLS}$	1.04	0.54	0.49	–0.02
$\widehat{\alpha}_t^{1,2SLS}$	0.99	0.49	0.44	–0.06
$\widehat{\beta}_t^{1,2SLS}$	1.00	1.00	0.45	0.44
$\widehat{\gamma}_t^{1,2SLS}$	0.99	0.49	0.44	–0.06
Exclusion Restrictions				
D1: $\mathbb{E}[Z_{ti}^T E_{ti}]$	0.00	0.00	–0.09	–0.09
D2: $CORR(Y_{ti},\, Z_{ti}^T)\|D_{ti}, E_{ti}^P$	0.00	0.00	0.00	0.00
Selection into Covariates				
$\mathbb{E}[X_{ti}\|D_{ti} > 0]$	0.00	0.36	0.00	0.36
$\mathbb{E}[X_{ti}\|D_{ti} < 0]$	0.00	0.64	0.00	0.64
$\mathbb{E}[E_{ti}\|D_{ti} > 0]$	0.01	0.01	0.42	0.42
$\mathbb{E}[E_{ti}\|D_{ti} < 0]$	–0.01	–0.01	0.71	0.71

Selection into E:

E_{ti}^P is Permanent, E_{ti}^M is Malleable

$E_{ti} \leftleftarrows 0.25E_{ti}^P + 0.75E_{ti}^M$

$E_{ti}^P \sim iidU[-\frac{1}{2}, \frac{1}{2}]$

$E_{ti}^M \leftleftarrows 1 - D_{ti}$ if $U_{ti}^E \leq 0.75$

$E_{ti}^M \sim iidU[-\frac{1}{2}, \frac{1}{2}]$ if $U_{ti}^E > 0.75$

$U_{ti}^E \sim U[0, 1]$

Selection into D:

$D_{ti} \leftleftarrows 0.5Z_{ti}^T + 0.25U_{ti}^D + 0.25E_{ti}^P$

$Z_{ti}^T, U_{ti}^D \sim iidU[-1, 1]$

Note: The specified DGPs were used to generate 100,000 observations for the previous and current time periods. The precise functions f_t^X and f_t^E used in each simulated DGP are specified in the text.