STRUCTURAL INSTRUMENTAL VARIABLE ESTIMATOR FOR DYNAMIC TREATMENT EFFECTS: SPANKING EFFECT ON BEHAVIOR

(February 13, 2006)

Myoung-jae Lee*
Department of Economics
Korea University
Anam-dong, Sungbuk-gu
Seoul 136-701, South Korea
myoungjae@korea.ac.kr

Fali Huang
School of Economics and Social Sciences
Singapore Management University
90 Stamford Road
Singapore 178903
flhuang@smu.edu.sg

Finding the effects of multiple sequential treatments on a response variable measured at the end of a trial is difficult, if some treatments are affected by interim responses; e.g., assessing the effects of sequential spankings on child behavior when parents adjust their interim spanking levels depending on interim behaviors. A headway, '*G estimation*', has been made in 1980's generalizing the usual static single-treatment effect analysis under 'no unobservable confounders' or 'selection on observables'. But G estimation is not easy to implement. In this paper, firstly, we propose a simpler alternative to G estimation—*instrumental variable estimator (IVE) for a linear structural model*—and show that our proposal and G estimation identify the same effect under some assumptions. Secondly, we explore the relation between our proposal and *Granger causality* to show that our approach is more general, although the two become equivalent for testing non-causality under a stationarity-type assumption. Thirdly, our approach and G estimation are applied to find the effects of spanking on child behavior. We find that mild spanking at early years reduces child behavior problems later, which seems to differ from most findings in the psychology/education literature.

Key words: dynamic model, treatment effect, panel data, instrumental variables, causality, spanking

* Corresponding Author

1 Introduction

Finding the effect of a treatment on a response variable is something relevant to all disciplines of science. The 'potential response' causal framework for treatment effect analysis with 'static one shot' binary treatment is by now well known (e.g., Rubin 1974, 2005, Holland 1986, Rosenbaum 2002, and Lee 2005). For a treatment d_i taking on $0, 1$ for individual $i = 1, ..., N$, two potential responses y_i^1 (treated) and y_i^0 (untreated) are envisioned. The individual treatment effect is $y_i^1 - y_i^0$, which is, however, never identified. Instead, the mean effect $E(y^1 - y^0)$ is estimated usually, although other effects such as $Median(y^1 - y^0)$ can be also of interest (Lee 2000).

Sometimes, multiple treatments $d = (d_1, ..., d_T)'$ are administered over time, and finding the *total dynamic treatment effect of the treatment (profile) d on a response variable measured at the end of a trial* is necessary. At first glance, it seems possible to strip the time dimension off to handle the multiple treatments as a single multinomial treatment. In this "naive" approach, one may set up and estimate a model for the final response as a (linear) function of all treatments (and some covariates) as if the treatments were given at the same time.

But a fundamental problem occurs to this naive approach when interim treatments are affected by interim responses, as this makes the time dimension indispensable and the interim treatments endogenous. For this case, Robins (1986) introduced 'G estimation'—or 'G (computation) algorithm'—which has been further refined/extended as can be seen in Robins (1998, 1999). But G estimation is not so easy to implement and does not show what the total effect consists of. The goal of this paper is to introduce a simpler linear-model-based alternative to G estimation and apply this method to an important educational question: does spanking work? Since spanking is done many times over time to improve child behavior, and since the parents adjust interim spankings based on interim responses (i.e., interim child behaviors), spanking/behavior nexus fits well our dynamic causal framework. The proposed method, which uses instrumental variable estimator (IVE), is easier to apply and more informative by decomposing the total effect into various direct and indirect effects.

The rest of this paper is organized as follows. Section 2 introduces a two-period model—the simplest dynamic setup—to present the desired (total dynamic) treatment effect and to demonstrate that the above naive approach fails to identify the treatment effect. Section 3 shows that our proposal can identify the desired effect and then compares our proposal to

'Granger causality', which is a probabilistic causality concept widely used. Section 4 reviews the original G estimation and a closely related approach, 'structural nested model', and then provides the details on how to implement our proposal; thus all methodologies are laid out in Section 4. Section 5 introduces our data drawn from the National Longitudinal Survey of Youth (NLSY). Section 6 presents empirical findings using our method, G estimation, the structural nested model, and Granger causality. Finally, Section 7 concludes. Throughout, we will use the two-period model, which our empirical analysis fits; the appendix contains a three-period extension and proofs.

2 Failure of Naive Dynamic Panel Data Model

Suppose what is observed is

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix},\ (d_1, \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}),\ (d_2, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix})$$

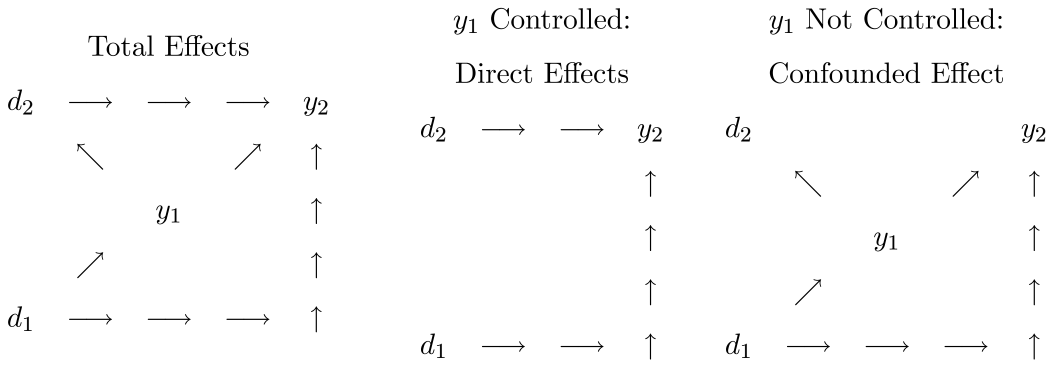
where x_0 and y_0 are the baseline covariate and response, and treatment d_t at period t temporally precedes $(x_t', y_t)'$, $t = 1, 2$. Here the subscript $i = 1, ..., N$ indexing individuals as in d_{it} is omitted; i will be also often omitted in the remainder of this paper. In our empirical analysis later with $T = 3$ periods, to assure that d_t precedes $(x_t', y_t)'$, the treatment at $t-1$ is used as d_t, which means that the original d_3 was discarded, leaving only two treatments d_1 and d_2.

Define

$$y_2^{jk} : \text{potential response at } t = 2 \text{ when } d_1 = j \text{ and } d_2 = k \text{ are exogenously set;}$$

also define y_1^j as the potential response at $t = 1$ when $d_1 = j$. The observed responses are $y_1 = y_1^{d_1}$ and $y_2 = y_2^{d_1 d_2}$. The expression 'exogenously set' has been also called 'intervened (by the authority)' in the literature as opposed to 'self-selected (by the individual)'. Pearl (2000) calls the former 'actively done (by the authority)' and the latter 'passively observed (to the statistician)'. Our goal is to find the mean effect $E(y_2^{jk} - y_2^{00})$ of the treatment profile $d = (d_1, d_2)'$ taking (j, k) versus no treatment $(0, 0)$.

To appreciate the aforementioned fundamental problem in dynamic treatment effect analysis, examine the graph 'Total Effects' omitting y_0, x_0, x_1, x_2, where d_2 has only a direct effect on y_2 whereas d_1 has direct *and* indirect (through y_1) effects; the desired $E(y_2^{jk} - y_2^{00})$ contains all these effects. The main difficulty is that d_2 is affected by the interim response y_1, which leads to the following dilemma. If y_1 is controlled (i.e., fixed), then the indirect effect of d_1 on y_2 is not identified as in the graph 'y_1 Controlled'. If y_1 is not controlled, then the effect of d_2 on y_2 gets distorted as y_1 becomes a 'common factor' (confounder) for d_2 and y_2: even if there is no effect of d_2 on y_2, a spurious effect may be found due to not controlling y_1 as in the graph 'y_1 Not Controlled'.



Total Effects

y_1 Controlled: Direct Effects

y_1 Not Controlled: Confounded Effect

The total (direct+indirect) effect composition with a confounder (y_1) can occur also with a covariate $w_1 \neq y_1$. For instance, y_2 can be death, w_1 is a 'death-predictor', and d_1 and d_2 are some medical procedures; just imagine replacing y_1 with w_1 in the graph 'Total Effects'.

Facing the multiple treatments $(d_1, d_2)'$, an applied researcher might be tempted to use a *'first-lag' dynamic model*

$$y_{i2} = \beta_1 + \beta_y y_{i1} + \beta_{d1} d_{i1} + \beta_{d2} d_{i2} + \beta'_{x2} x_{i2} + v_{i2}$$

where β_1, β_y, β_{d1}, β_{d2} and β_{x2} are parameters and v_{i2} is an error term. But this model fails to identify the desired total effect, because the *indirect effect of d_1 on y_2 through y_1 is missed* by controlling y_1. Intuitively, if the effect of d_1 on y_1 is γ_d, then the indirect effect of d_1 on y_2 through y_1 is $\beta_y \gamma_d$. The first-lag dynamic model can identify only the direct effects β_{d1} and β_{d2} of d_1 and d_2 on y_2, whereas the desired (total dynamic treatment) effect is the sum of

$$\begin{aligned} &\textit{direct and indirect effects of } d_1 \text{ on } y_2 &:& \quad \beta_{d1} + \beta_y \gamma_d, \\ &\textit{direct effect of } d_2 \text{ on } y_2 &:& \quad \beta_{d2}. \end{aligned}$$

To see exactly how $E(y_2^{jk} - y_2^{00})$ contains all the direct and indirect effects, we need to have the β and γ parameters in the model for y_2^{jk} (then the y_2 model will be derived from the y_2^{jk} model). For this, suppose

$$\begin{aligned} y_{i1}^{j} &= \gamma_1 + \gamma_y y_{i0} + \gamma_d j + \gamma_{x1}' x_{i1} + v_{i1}, \\ y_{i2}^{jk} &= \beta_1 + \beta_y y_{i1}^{j} + \beta_{d1} j + \beta_{d2} k + \beta_{x2}' x_{i2} + v_{i2} \end{aligned} \tag{2.1}$$

where $\gamma_1, \gamma_y, \gamma_d, \gamma_{x1}$ are the parameters for the y_{i1}^j model and v_{i1} is an error term. Substitute out y_{i1}^j to get

$$y_2^{jk} = (\beta_1 + \beta_y\gamma_1) + \beta_y\gamma_y y_0 \ + (\beta_{d1} + \beta_y\gamma_d)j + \beta_{d2}k \ + \beta_y\gamma_{x1}'x_1 + \beta_{x2}'x_2 \ + (\beta_y v_1 + v_2). \tag{2.2}$$

From this,

$$\begin{aligned} y_2^{jk} - y_2^{00} = (\beta_{d1} + \beta_y\gamma_d)j + \beta_{d2}k = E(y_2^{jk} - y_2^{00}) \\ \Longrightarrow \quad E(y_2^{jk} - y_2^{00}) = (\beta_{d1} + \beta_y\gamma_d) + \beta_{d2} \quad \text{when } j = k = 1. \end{aligned} \tag{2.3}$$

Although we allow non-binary treatments for more generality, often we will use binary treatments to convey the essential ideas as done just now.

The model (2.1) can be generalized in many ways. Just to list a few, first, the treatments may interact among themselves as well as with covariates. Second, the parameters may be time-dependent. Third, nonlinear functions of the treatments may be more appropriate. Fourth, not just contemporaneous covariates, but also lagged covariates may matter for the current potential response. Fifth, v_{i1}^j and v_{i2}^{jk} instead of v_{i1} and v_{i2} may appear. All of these can be accommodated without much difficulty. For instance, the last generalization would entail

$$y_2^{jk} - y_2^{00} = (\beta_{d1} + \beta_y\gamma_d)j + \beta_{d2}k + (\beta_y v_1^j + v_2^{jk}) - (\beta_y v_1^0 + v_2^{00}).$$

With the usual zero-mean assumption for the error terms, $(\beta_y v_1^j + v_2^{jk}) - (\beta_y v_1^0 + v_2^{00})$ disappears in $E(y_2^{jk} - y_2^{00})$. To ease exposition, however, we will stick to the basic model (2.1) unless otherwise necessary.

3 Total Effect Identified by Linear Structural Model

Although the desired effect $E(y_2^{jk} - y_2^{00})$ was defined in terms of potential responses, estimation can be done only with the observed responses. Hence we need to derive y_1 and y_2 equations from their potential response equations. Comparing (2.1) to the y_2 equation (the first-lag model) that appeared before (2.1), one may get the impression that the y_1 and y_2

equations can be obtained simply by replacing j and k in the y_1^j and y_2^{jk} equations with d_1 and d_2. But the replacement is in fact equivalent to a '*no unobserved confounder (NUC)*' condition—also called 'selection on observables'—as follows.

Suppose

$$E(y_1^j|y_0, x_1) = g(j, y_0, x_1) \quad \text{and} \quad E(y_2^{jk}|y_1, x_2) = h(j, k, y_1, x_2) \tag{3.1}$$

for a function g and h and the covariates x_1 and x_2; g and h include the linear functions in (2.1) as special cases. Under this, the appendix shows

$$\begin{aligned} \text{NUC (i)} \quad &: \quad E(y_1^j|d_1 = j, y_0, x_1) = E(y_1^j|y_0, x_1) \Longleftrightarrow E(y_1|d_1, y_0, x_1) = g(d_1, y_0, x_1) \\ \text{NUC (ii)} \quad &: \quad E(y_2^{jk}|d_1 = j, d_2 = k, y_1, x_2) = E(y_2^{jk}|y_1, x_2) \\ &\Longleftrightarrow \quad E(y_2|d_1, d_2, y_1, x_2) = h(d_1, d_2, y_1, x_2). \end{aligned} \tag{3.2}$$

The left-hand sides of '$\Longleftrightarrow$' are the NUC that d_1 *and* d_2 *are as good as randomized for* y_1^j *and* y_2^{jk}, *conditional on the observables.* The right-hand sides are replacing the fixed (i.e., intervened) j and k with random (i.e., self-selected) d_1 and d_2. Note that the observables for y_2 include y_1, which allows y_1 affecting d_2. NUC will appear again for G estimation in a different form.

In view of (3.2), setting g and h as the linear functions in (2.1) and adopting NUC (i) and (ii), we get the equations for y_1 and y_2:

$$\begin{aligned} y_{i1} &= \gamma_1 + \gamma_y y_{i0} + \gamma_d d_1 + \gamma_{x1}' x_{i1} + v_{i1}, \\ y_{i2} &= \beta_1 + \beta_y y_{i1} + \beta_{d1} d_1 + \beta_{d2} d_2 + \beta_{x2}' x_{i2} + v_{i2}. \end{aligned} \tag{3.3}$$

Using these, we can estimate the γ and β parameters, with which the desired effect is estimated. This is a two-step method, as two equations are estimated. How to estimate (3.3) specifically is data-dependent and will be discussed later.

Suppose

$$\textit{equal contemporaneous effects} : \gamma_d = \beta_{d2} \tag{3.4}$$

that the effect of d_1 on y_1 is the same as the effect of d_2 on y_2. This is a stationarity-type assumption, under which

$$d_1 \text{ effect is } \beta_{d1} + \beta_y \beta_{d2} \quad \text{and} \quad d_2 \text{ effect is } \beta_{d2}.$$

These are identified with the y_2-equation only. Hence under (3.4), it is not necessary to estimate the y_1 equation, which can be helpful if estimating the y_1 equation is difficult.

Instead of estimating only the y_2 equation under $\gamma_d = \beta_{d2}$, another way to dispense with the y_1 equation is substituting the y_1-equation into the y_2 equation in (3.3) to get the observed version of (2.2):

$$y_2 = (\beta_1 + \beta_y \gamma_1) + \beta_y \gamma_y y_0 + (\beta_{d1} + \beta_y \gamma_d) d_1 + \beta_{d2} d_2 + \beta_y \gamma'_{x1} x_1 + \beta'_{x2} x_2 + (\beta_y v_1 + v_2). \quad (3.5)$$

The NUC from (2.2) to (3.5) is

$$\text{NUC (iii)}: \; E(y_2^{jk} | d_1 = j, d_2 = k, y_0, x_1, x_2) = E(y_2^{jk} | y_0, x_1, x_2)$$

which differs slightly from NUC (ii). Model (3.5) for y_2 is unusual in that the last-lagged y_0, not the first-lagged y_1, is included. The total effect (2.3) can be identified with (3.5) as the sum of the coefficients of d_1 and d_2. This 'last-lag model' is simpler than the 'first-lag' model (3.3), but there is a disadvantage: the decomposition of the total effect of d_1 cannot be done. Another disadvantage is that (3.5) could be harder to estimate than the first-lag model in (3.3) as to be discussed later.

For our empirical analysis on spanking later, here we note one extension of (3.3). Even if spanking is beneficial, too much spanking is likely to be harmful: i.e., spanking effects could be nonlinear. Suppose that the effects of d_1 and d_2 are quadratic to yield

$$\begin{aligned} y_1 &= \gamma_1 + \gamma_y y_0 + \gamma_d d_1 + \gamma_{dq} d_1^2 + \gamma'_{x1} x_1 + v_1, \\ y_2 &= \beta_1 + \beta_y y_1 + \beta_{d1} d_1 + \beta_{d1q} d_1^2 + \beta_{d2} d_2 + \beta_{d2q} d_2^2 + \beta'_{x2} x_2 + v_2. \end{aligned} \quad (3.6)$$

With the first derivatives, the three key effects are

$$\begin{aligned} \text{direct and indirect effects of } d_1 &= j: \; \beta_{d1} + 2\beta_{d1q} j, \;\; \beta_y(\gamma_d + 2\gamma_{dq} j) \quad (3.7) \\ \text{direct effect of } d_2 &= k: \; \beta_{d2} + 2\beta_{d2q} k. \end{aligned}$$

Instead of (3.6), one may use its last-lag version with y_1 substituted out, which is analogous to (3.5).

For two time-series $\{y_t\}$ and $\{d_t\}$, 'Granger non-causality' (Granger 1969, 1980) is often tested by $H_0 : \beta_{d1} = \beta_{d2} = 0$ in

$$y_2 = \beta_1 + \beta_{y1} y_1 + \beta_{y0} y_0 + \beta_{d1} d_1 + \beta_{d2} d_2 + \text{covariates} + v_2 \quad (3.8)$$

where *all lagged* d and y appear on the right-hand side. But this *Granger non-causality test is only for the direct effect of* d, because y_1 is included in the right-hand side. Interestingly, this problem disappears under (3.4) $\gamma_d = \beta_{d2}$, because the indirect effect $\beta_y \gamma_d$ becomes zero when the two direct effects β_{d1} and β_{d2} are zero. This solution, however, works only for *the test* of non-causality. For the effect magnitude, (3.8) still misses the indirect effect. These points were in essence noted by Robins et al. (1999).

4 Three Estimation Methods

So far we showed that the usual first-lag dynamic model and the Granger causality model fail to identify the desired effect, which the linear models such as (3.3) or (3.6) can identify. This section briefly reviews G estimation and a simple version of structural nested model. Also we provide the details on how to estimate the linear models in practice. Thus, all three methodologies to be used for our empirical analysis later are laid out in this section together.

4.1 G Estimation

Define

$$X_2 \equiv (x_0', x_1', x_2')',$$

and '$a \amalg b|c$' as the conditional independence of a and b given c.. Assume the following form of NUC or 'selection-on-observables (y_0, X_2)':

$$\text{NUC (a):}\quad y_2^{jk} \amalg d_1 \;|(y_0, X_2) \qquad\qquad \text{NUC (b):}\quad y_2^{jk} \amalg d_2 \;|(d_1, y_1, y_0, X_2). \tag{4.1}$$

NUC (i) in (3.2) is for y_1^j, whereas NUC (a) and (b) are only for y_2^{jk}. Due to the different conditioning sets, NUC (ii)-(iii) are neither weaker nor stronger than NUC (a) and (b), despite some similarities.

With $f(y_1|d_1, y_0, X_2)$ denoting the conditional density/probability for $y_1|(d_1, y_0, X_2)$, G estimation under NUC (a) and (b) is

$$E(y_2^{jk}|y_0, X_2) = \int E(y_2|d_1 = j, d_2 = k, y_1, y_0, X_2) f(y_1|d_1 = j, y_0, X_2)\partial y_1 \tag{4.2}$$

where '∂' is used instead of 'd' to prevent confusion with treatment d. The right-hand side is identified, and so is the conditional mean $E(y_2^{jk}|y_0, X_2)$. Then

$$E(y_2^{jk} - y_2^{00}) = \int \{E(y_2^{jk}|y_0, X_2) - E(y_2^{00}|y_0, X_2)\}\partial F(y_0, X_2). \tag{4.3}$$

where $F(y_0, X_2)$ denotes the distribution of (y_0, X_2). The appendix shows that *(4.3) equals (2.3) for the linear models.*

The G estimation works because the right-hand side of (4.2) is

$$\begin{aligned} & \int E(y_2^{jk}|d_1 = j, d_2 = k,\ y_1, y_0, X_2) f(y_1|d_1 = j, y_0, X_2)\partial y_1 \\ = & \int E(y_2^{jk}|d_1 = j,\ y_1, y_0, X_2) f(y_1|d_1 = j, y_0, X_2)\partial y_1 \qquad \text{due to NUC (b)} \\ = & \ E(y_2^{jk}|d_1 = j, y_0, X_2) \qquad \text{for } y_1 \text{ is integrated out} \\ = & \ E(y_2^{jk}|y_0, X_2) \qquad \text{due to NUC (a).} \end{aligned}$$

The last line shows that NUC (a) is non-essential, because we may stop at the second-from-the-last line to identify only $E(y_2^{jk}|d_1 = j, y_0, X_2)$. Then $E(y_2^{jk} - y_2^{00}|d_1 = j)$—the effect on the treated $d_1 = j$—can be obtained by doing analogously to (4.3).

To implement (4.2), one has to find the conditional mean and the conditional density, and then integrate out y_1. This is rather difficult. But (4.2) gets simplified much if y_1 is binary. With a binary y_1, (4.2) becomes

$$\begin{aligned} E(y_2^{jk}|y_0, X_2) &= P(y_2 = 1|d_1 = j, d_2 = k, y_1 = 0, y_0, X_2) \cdot P(y_1 = 0|d_1 = j, y_0, X_2) \\ +P(y_2 &= 1|d_1 = j, d_2 = k, y_1 = 1, y_0, X_2) P(y_1 = 1|d_1 = j, y_0, X_2). \qquad (4.4) \end{aligned}$$

Apply probit (or logit) to y_2 on d_1, d_2, y_1, y_0, X_2 to obtain the two probit probabilities for $y_2 = 1$:

$$\Phi(\psi_1 + \psi_{d1} d_1 + \psi_{d2} d_2 + \psi_{y1} y_1 + \psi_{y0} y_0 + \psi_x' X_2).$$

Also apply probit to y_1 on d_1, y_0, X_2 to get the probit probabilities for $y_1 = 0, 1$:

$$\Phi(\eta_1 + \eta_{d1} d_1 + \eta_{y0} y_0 + \eta_x' X_2).$$

Substituting these into (4.4), it is straightforward to get $E(y_2^{jk}|y_0, X_2)$. This binary-y_1 version will be used for our empirical analysis later.

4.2 Structural Nested Model

Instead of G estimation, there are other estimation methods available for dynamic causal inference (Robins 1998, 1999). A particularly easy method to apply is the following version in Robins (1992) of 'structural nested model', which will be also applied to our data. An epidemiological application can be seen in Witteman et al. (1998) among others.

Suppose, for an unknown parameter ψ_o,

$$y_2^{00} = y_2^{jk}\frac{\exp(\psi_o j)+\exp(\psi_o k)}{2} \iff y_2^{jk} = y_2^{00}\frac{2}{\exp(\psi_o j)+\exp(\psi_o k)}. \tag{4.5}$$

The treatments multiplicatively alter the no-treatment response y_2^{00}. The mean effect $E(y_2^{jk} - y_2^{00})$ is

$$E[y_2^{00}\cdot\{\frac{2}{\exp(\psi_o j)+\exp(\psi_o k)} - 1\}].$$

If $y_2^{jk} \amalg d_2|observables$ as in NUC(b), then (4.5) implies $y_2^{00} \amalg d_2|observables$ as well. Defining

$$S_i(\psi) \equiv y_{i2}\frac{\exp(\psi d_{i1})+\exp(\psi d_{i2})}{2},$$

we get $S_i(\psi_o) = y_{i2}^{00}$. Transforming the treatments into binary treatments, the true value of θ in the following logit should be zero if $\psi = \psi_o$: for some parameter β_2,

$$P(d_2 = 1|y_1, y_0, d_1, X_2) = \frac{\exp\{\beta_2'(y_1, y_0, d_1, X_2') + \theta S(\psi)\}}{1+\exp\{\beta_2'(y_1, y_0, d_1, X_2') + \theta S(\psi)\}}.$$

Depending on ψ, we get different asymptotic t-values $t_N(\psi)$ for θ. Following the duality between a test and the confidence interval (CI), a 95% CI for ψ is $\{\psi : |t_N(\psi)| < 1.96\}$. The middle point of the CI or ψ for $\hat{\theta} = 0$ may be used as a point estimator $\hat{\psi}$ of ψ where $\hat{\theta}$ is the logit estimator for θ. The main disadvantage of this approach is the same-effect restriction for d_1 and d_2 and the arbitrary functional form assumption linking all counter-factuals y_2^{jk} to y_2^{00}, but the main advantage—straightforward to implement—seems incomparable with other dynamic causal effect estimators.

The same-effect assumption can be relaxed: adopt for some parameters ψ_0, ψ_1, θ_1, and θ_2,

$$S_2(\psi_0, \psi_1) \equiv y_2\frac{\exp(\psi_0 d_1)+\exp(\psi_1 d_2)}{2} \quad \text{and}$$

$$P(d_2 \quad = \quad 1|y_1, y_0, d_1, X_2) = \frac{\exp\{\beta_2'(y_1, y_0, d_1, X_2') + \theta_1 S_2(\psi_0,\psi_1) + \theta_2 S_2(\psi_0,\psi_1)^2\}}{1+\exp\{\beta_2'(y_1, y_0, d_1, X_2') + \theta_1 S_2(\psi_0,\psi_1) + \theta_2 S_2(\psi_0,\psi_1)^2\}}.$$

We well set $\psi_1 = c\psi_0$ in our empirical analysis later to estimate ψ_0 from each fixed level of c. As c changes around one, the estimate for ψ_0 will change, showing how robust the empirical results are to the same-effect assumption.

4.3 IVE for Structural Linear Models

Although IVE is no stranger to the treatment effect literature as can be seen in Angrist and Imbens (1995), Angrist, et al. (1996), and Imbens and Rubin (1997), here we briefly describe IVE. For a linear model $y_i = x_i'\beta + u_i$ with $E(xu) \neq 0$, least squares estimator (LSE) is not applicable. But suppose there is an 'instrument' z for x such that $E(zx')$ is of full column rank and $E(zu) = 0$. IVE in its wide sense is any method-of-moment estimator based on $E(zu) = 0$, and IVE b_{ive} in its narrow sense takes the form

$$\begin{aligned} b_{ive} &= \{\sum_i x_i z_i' (\sum_i z_i z_i')^{-1} \sum_i z_i x_i'\}^{-1} \cdot \sum_i x_i z_i' (\sum_i z_i z_i')^{-1} \sum_i z_i y_i; \\ \sqrt{N}(b_{ive} - \beta) &\rightsquigarrow N\{0,\ GE(zz'u^2)G'\} \quad (\text{`} \rightsquigarrow \text{' for convergence in law) where} \\ G &\equiv [E(xz')\{E(zz')\}^{-1}E(zx')]^{-1}E(xz')\{E(zz')\}^{-1}. \end{aligned}$$

A more efficient version is available ('generalized method-of-moment' or 'estimating-equation' estimator), but we will use the IVE to simplify exposition; the IVE is also often said to have a better finite sample performance than the efficient version. With IVE (or LSE), we can estimate the γ and β parameters in (3.3) and thus the dynamic treatment effect.

The main practical question for IVE is where to find instruments. In panel data, this task depends on the source of endogeneity. Recall the y_2 equation in (3.3):

$$y_2 = \beta_1 + \beta_y y_1 + \beta_{d1} d_1 + \beta_{d2} d_2 + \beta_{x2}' x_2 + v_2 \qquad (4.6)$$

with all regressors (y_1, d_1, d_2, x_2')' potentially endogenous. Suppose that v_{it} consists of a time-constant error δ_i and a time-variant error u_{it}, and that x_{it} is correlated with δ_i but not with u_{it} at all leads and lags:

$$v_{it} = \delta_i + u_{it}, \quad COR(\delta_i, x_{it}) \neq 0 \ \forall t, \quad COR(u_{is}, x_{it}) = 0 \ \forall s, t. \qquad (4.7)$$

In this case, first-difference the model to get

$$y_2 - y_1 = \beta_1 - \gamma_1 + \beta_y y_1 - \gamma_y y_0 + (\beta_{d1} - \gamma_d) d_1 + \beta_{d2} d_2 + \beta_{x2}' x_2 - \gamma_{x1}' x_1 + u_2 - u_1. \quad (4.8)$$

Some elements of (y_0, x_0', x_1', x_2') may be used as instruments for the regressors $(y_1, y_0, d_1, d_2, x_2', x_1')'$ in (4.8).

Instead of (4.7), the subject may adjust their current and future x_{it} after observing the current u_{it}. In this case, it may hold that

$$COR(u_{is}, x_{it}) \neq 0 \ \forall s \leq t \quad \text{but} \quad COR(u_{is}, x_{it}) = 0 \ \forall s > t. \qquad (4.9)$$

More on finding moment conditions in panel data can be seen in Lee (2002). Certainly, assumptions weaker/stronger than (4.7) and (4.9) may be invoked. As these examples illustrate, typically lagged responses and covariates are the sources for instruments, and this aspect explains why it might be difficult to estimate the last-lag model (3.5): the first-differenced model of (3.5) includes all period covariates. The instruments used for our data will be explained later.

One may get the impression that IVE is rather arbitrary. But G estimation also includes (even stronger) moment conditions. To appreciate this point, suppose

$$d_1 = \alpha_{11} + \alpha'_{1x} x_0 + \alpha_{1y} y_0 + \varepsilon_1 \quad \text{and} \quad y_2^{jk} = \beta_1 + \beta_y y_1^j + \beta_{d1} j + \beta_{d2} k + \beta'_{x2} x_2 + v_2 \text{ in (2.1)}$$

for some parameters $(\alpha_{11}, \alpha'_{1x}, \alpha_{1y})$ and an error term ε_1. NUC (a) requires $v_2 \amalg \varepsilon_1 | (y_0, X_2)$, under which d_1 becomes exogenous to v_2 in (4.6). With IVE, we try to replace this kind of assumptions with weaker ones—e.g., $E(x_0 v_2) = 0$ instead of $E(d_1 v_2) = 0$ to allow d_1 to be endogenous in the y_2 equation. Roughly speaking, applying LSE to the linear models in (3.3) would be analogous to G estimation, whereas IVE is an "elaborate" attempt, by taking advantage of the linear model assumption, to weaken strong assumptions such as the observables-conditional treatment exogeneity

5 Data

The NLSY79 child sample contains rich information on children born to women respondents of the NLSY79. Starting from 1986, questionnaires were developed to collect information about the cognitive, social, and behavioral development of the children, as well as their family backgrounds and detailed home inputs. The surveys were conducted every two years, which enables us to get detailed information when the children were 2-3, 4-5, and 6-7 years old. Based on the surveys during 1986-1998, we constructed a longitudinal sample of 1329 children. The summary statistics of all variables in this sample are listed in Table 0.

The survey question on spanking asks the mother: "*About how many times, if any, have you had to spank your child in the past week?*" Spanking is quite common for young kids, although its frequency decreases as a child grows: 87% mothers spanked their toddlers at least once in the past week, while only 68% spanked their five-year-olds. Since most children are spanked modestly in frequency, our study will focus on the effects of modest spanking that are debated most (Baumrind et al. 2002). This led to a sample of 961 children spanked up to three times a week before age three (73% of the whole sample) and up to five times a week before age five (94% of the whole sample). It is our main working sample on which most of our empirical analyses are based. Because all children in the sample were spanked not more than several times a week, and also because the answered spanking frequency may not be representative of the "regular" spanking frequency—the questionnaire is only for the past week—we also use a binary variable for spanking (1 if ever spanked and 0 otherwise). In most cases, estimation results using both the original spanking frequency and its binary version are presented.

The social and behavioral development of children above four years old is measured by the *Behavior Problems Index (BPI).* BPI is one of the most frequently used variables in the NLSY79 child assessments for a wide range of child attitude and behavior. *A higher BPI represents more behavior problems.* In a fully representative sample of children, the mean standard score is expected to be 100. The BPI in our sample has mean 105.3 and standard deviation (SD) 14.7 around age 6-7, and mean 104.8 with SD 14.8 around age 4-5. Two binary variables are also constructed for BPI: 1 if BPI is higher than the sample mean and 0 otherwise. Since there is no BPI for age below four, we use two sets of health assessments to measure a child's initial behavior problems. Motors and Social Development Scale (MSD) measures developmental milestones in the areas of motor, cognitive, communication, and social development; Temperament Scales measure temperament or behavioral style, where the most relevant scales for children before age three are measures of their compliance, attachment and sociability.

The link between spanking and behavior problems seems to be a complicated one, as it is still hotly debated after many years of investigation (Gershoff 2002). The difficulty in establishing the causal link is the endogeneity of spanking arising from various sources. For example, inappropriate home inputs may induce poor behaviors of children and more spanking from the parents; if the detailed home inputs are not properly controlled, then the positive correlation between spanking and behavior problems may be spurious. As shown in Table 1, children with better home environments are indeed spanked less and have fewer behavioral problems. For instance, mothers who often read to their children at age 2-3 were less likely to spank them than those that did not; their children had better early development results and fewer behavior problems later. Similar patterns hold for children who have more books and less TV hours at home, and who were breast-fed. This regularity is still true when the overall home environment is measured by HOME (a simple summation of the individual input scores), which is, however, much less informative than the detailed individual inputs.

The strength of our data is that, in addition to key family background variables, there is a rich set of home inputs containing about twenty variables at both age 6-7 and 4-5, and ten variables at age 2-3. This would greatly reduce potential omitted variable biases.

6 Empirical Results

6.1 IVE for Structural Linear Model

We first estimate the structural linear models as in (3.3), where y_2 and y_1 are BPI scores at age 6-7 and 4-5; d_2 and d_1 are the spanking frequencies at age 4-5 and 2-3, or their binary versions (ever spanking or not). Under the assumption that only the current home inputs are related to current BPI, we use past home inputs as instrumental variables for the current inputs; details on covariate choices when a covariate might be affected either by a treatment or by a response are provided in the appendix. The IVE/LSE results are in Table 2.

Column IV(B) for binary spanking uses (as covariates) the child race, sex, and birth order (in G1 of Table 0), the family background variables (G2 of Table 0), and the current home inputs at age 6-7 (G3) which are instrumented by the earlier home inputs at age 2-3 and 4-5 (G4 and G5). The effect of spanking at age 2-3 is insignificantly negative, while the effect of spanking at age 4-5 is insignificantly positive; the magnitude is, however, about 3 times greater in the former than in the latter. The lagged response y_1 has a significant positive effect with the effect magnitude 0.52.

In column IV(B'), the same specification is used except that the exact numbers of spanking and their squared terms are used instead of the binary spanking variables in IV(B). Recalling (3.7) and using column IV(B') in the right half of Table 2 for BPI at age 4-5, we get

$$\begin{aligned} &\text{direct and indirect effect of } d_1: -3.11+1.24d_1 \quad \text{and} \quad 0.48(-4.11+2.04d_1) \\ \Longrightarrow\quad &\text{total effect of } d_1: -5.08+2.22d_1; \\ &\text{effect of } d_2: -1.60+1.56d_2. \end{aligned} \tag{6.1}$$

Moderate spanking reduces BPI as the negative 'intercepts' indicate, but the effect turns harmful as d_1 (d_2) takes 3 (2) as the positive 'slopes' show. The magnitudes are greater for d_1. This result is coherent with the preceding paragraph—still statistically insignificant though.

One problem with the models for IV(B) and IV(B') is that the instruments are weak with low correlations with the endogenous variables. Hence exploring ways to avoid IVE, we classify the home inputs into two groups: disciplinary measures and the others, because endogeneity problem is far more worrisome for the disciplinary measures than the others. The disciplinary inputs such as grounding children may be simultaneously related to BPI; i.e., BPI and disciplinary inputs may exchange influences. Imagine that spanking is done to influence BPI, which then affects the disciplinary inputs, which further affects BPI, and so on. In this case, as explained in the appendix, dropping the disciplinary inputs from the y_2 equation captures the full impact of spanking on BPI whereas keeping them captures only the initial effect of spanking with the disciplinary inputs held constant. This motivates removing the disciplinary inputs and then applying LSE. The result is shown in column LSE(BD) where two aggregate home environment scores HOME at age 2-3 and 4-5 are used as further controls. The effect of d_1 is now significantly negative and the effect of d_2 is insignificantly negative; also the magnitude for d_1 is much greater. The effect of y_1 is 0.44—about the same as those in IV(B) and IV(B').

The next four columns present the results for the y_1 equation, where the specification and estimation is the same as for the y_2 equation except that 'T' in IV(BT) and LSE(BTD) means that the three measures of child temperament (in G1 of Table 0) are used instead of MSD as proxies for y_0. The sample size decreases due to missing values in the temperament measures. The estimation results are similar to those for the y_2 equation. The only statistically significant finding is the large effect -14.7 of d_1 in column LSE(BTD).

Although we computed the total effect in (6.1), it was for the quadratic model using the exact spanking frequencies. As already mentioned, the exact frequencies may not be so reliable. Hence we compute the total effect as in (2.3) using the binary versions of d_1 and d_2; this effect will be used as the main effect of interest in this paper. First, from IV(B) for the y_2 equation, d_1 reduces y_2 by 4.03, while d_2 increases y_2 by 1.42; these are the direct effects, because y_1 is controlled. Second, the effect of y_1 on y_2 is 0.52. Third, from IV(B) for the y_1 equation, d_1 reduces y_1 by 3.60. So the indirect effect of d_1 on y_2 is $0.52 \times (-3.60) = -1.87$, which is about 46% of the direct effect 4.03. Hence, taken together, the total effect of d_1 is

$$\textit{direct effect} + \textit{indirect effect through } y_1 : \widehat{\beta}_{d1} + \widehat{\beta}_y \widehat{\gamma}_d = -4.03 + 0.52 \times (-3.60) = -5.90$$

which is 40% of SD(BPI). The bootstrap bias-corrected 95% CI is $(-21.1, 1.3)$, which is a nearly significant finding despite that only the effect of y_1 is significant. While d_1 seems to have the intended effect of reducing behavior problems, d_2 does not. The total effect of spanking $d = (d_1, d_2)'$ and its bootstrap bias-corrected 95% CI are, respectively,

$$-5.90\ (d_1 \text{ effect})\ +\ 1.42\ (d_2 \text{ effect}) = -4.48 \quad \text{and} \quad (-29.1, 11). \tag{6.2}$$

This CI includes 0, but is much longer on the negative side.

6.2 G Estimation

In order to apply the simplified G estimation with binary responses, we convert BPIs to dummy variables (higher than the sample mean or not). The binary spanking variables (ever spanked or not) are used as well to obtain the total effect with ease. The probit results are shown in Table 3, where the entries are the estimated marginal effects calculated at the sample means of the control variables. The probit is the discrete analog of the dynamic panel data model (but no unit-specific effect such as δ_i in (4.7) is considered in the probit), and as such, it misses the indirect effects; the desired total effect using (4.4) will be shown after the probit direct estimates are examined first. We also tried logit instead of probit, but the logit results are omitted, for they differ little from the probit results.

Column Probit(B) includes as controls the current and earlier home inputs as well as family background variables. Modest spanking at age 2-3 reduces the probability of higher-than-average BPI at age 6-7 by 0.35, which is significant at 10% level; in contrast, spanking at age 4-5 increases the same probability by 0.07, but it is insignificant. Column Probit reports results excluding the early inputs at age 2-3 and family background variables due to the endogeneity concern. The coefficient of spanking at age 2-3 is still negative and significant with p-value 5.6%, but its effect magnitude is reduced to -0.23; the explanatory power is also reduced in view of the pseudo R^2's, while the other results are very similar. The same trend continues in the third column where the disciplinary inputs at age 6-7 are taken out to further avoid the potential endogeneity problem. Overall, the general pattern is that binary d_1 reduces y_2, while binary d_2 tends to increase y_2. The latter effect, however, is not significant.

The probit results for y_1 are presented in the second part of the table. The first column includes the home inputs as well as family background variables; the second column excludes family backgrounds variables, and uses child temperament measures in addition to MSD to control for the child's initial characteristics; all of these variables are controlled in the last column. The coefficient of spanking is negative but insignificant in column Probit(B), while it becomes significant in the next two columns, the estimates of which suggest that binary d_1 reduces the probability of having higher-than-average BPI two years later by 0.44.

The desired total effect using (4.4) can be obtained with estimates in columns Probit and Probit(T) in Table 3 (other columns can be used as well, which makes little difference): the total effect of spanking at both age 2-3 and 4-5 is

$$\text{total effect} : E(y_2^{11}) - E(y_2^{00}) = 0.047; \ 95\% \text{ CI } (-0.4, 0.48).$$

With 0 sitting almost in the middle, this is not informative; a possible reason for this is that the control group with no spanking at age 2-3 is very small when relevant inputs are controlled. To see if anything can be learned from G estimation, we decompose the total effect into two parts: the effect of spanking at age 4-5 (jointly with spanking at age 2-3) and the effect of spanking at age 2-3 (jointly with no spanking at age 4-5), which are, respectively,

$$\begin{aligned} E(y_2^{11}) - E(y_2^{10}) &= 0.16; \ 95\% \text{ CI } (-0.14, 0.30) \\ E(y_2^{10}) - E(y_2^{00}) &= -0.12; \ 95\% \text{ CI } (-0.64, 0.35). \end{aligned}$$

Though both CI's still include zero, the estimates suggest that binary spanking at age 2-3 reduces the probability of having higher-than-average BPI at age 6-7, while binary spanking at age 4-5 tends to increase it. This pattern was noted also in the IVE results.

6.3 Nested Structural Model

The results for the structural nested model are in Table 4. The regressors in the first row include the detailed home inputs; measures of child temperament are added in the second row Logit(T), while family backgrounds variables are further added in the third row Logit(TB). The point estimate $\widehat{\psi_0}$ increases from 0 to 0.04 across the specifications as more controls are added, where $\widehat{\psi_0} = 0.04$ corresponds to 4.13 points reduction (about 28% reduction of one SD) of BPI at age 6-7. This magnitude does not differ much from those obtained using the IV methods above.

Since our earlier results suggest that the effects of spanking vary at different ages, we allow $\widehat{\psi_1} = c\widehat{\psi_0}$ under the specification of Logit(TB), where $\widehat{\psi_0}$ still indicates the effect of spanking at age 2-3, $\widehat{\psi_1}$ indicates the effect of spanking at age 4-5, and c is a positive number.

The estimated ψ_0 varies from 0.20 to 0.01 as c changes from 1/4 to 4, corresponding to a range of reduction 19.1 to 1.05 of BPI at age 6-7.

6.4 Granger Causality

Table 5 presents the results for the Granger causality model (3.8). The various specifications differ mainly in terms of the control variables used. With lagged BPI controlled, the lagged spanking is still significant, and thus *Granger non-causality is rejected.* In this case, as noted already, the coefficients of d_1 and d_2 show only their direct effects at best, which should be borne in mind in the following interpretation. The coefficients of spanking at age 2-3 are always negative and significant; the effects of spanking at age 4-5 are also negative, although mostly insignificant. Their magnitudes are similar to the IV estimates in Table 2. In the third column where the exact spanking frequencies are used, the effects of spanking are concave with significant estimates. The last column has the most comprehensive controls, including the current and earlier home inputs, child temperament measures as well as family background variables. The coefficients of the two spanking variables are both negative and significant.

7 Conclusions

In this paper, when a treatment is repeated over time and the final response is measured at the end, we showed how to estimate dynamic treatment effects with IVE and linear structural models. In our approach, interim treatments are allowed to have an immediate (direct) effect as well as a lingering (indirect) effect through interim responses; also, the interim treatments are allowed to be affected by interim responses. These feedbacks pose a dilemma to the usual dynamic model approach: if the interim responses are not controlled, then they become a confounder, because the treatment and control groups differ systematically in the interim responses; otherwise, the indirect effects are missed. An extreme form of this can be seen in the usual Granger causality model where all interim responses are controlled and consequently all indirect effects are missed. Nonetheless, we showed that, when the hypothesis of non-causality is not rejected, the Granger non-causality inference is valid under a stationary-type assumption. We also showed that our approach of IVE for linear structural models identifies the same total effect of the entire treatment 'profile' as the 'G estimation' of Robins (1986) does.

The IVE approach and two practical versions of G estimation were applied to find the effect of spanking on child behavior problems. The empirical results varied across different estimation methods, but coherently indicated that moderate spanking works, and spanking at early age of 2-3 has the stronger effect on reducing behavior problems at age 6-7 than spanking at age 4-5; these results seem consistent with the notion that spanking at earlier ages may stop a child's bad behaviors and hence prevent bad habits from forming in the beginning. Our preferred estimate suggested the overall effect (including direct and indirect effects) of spanking at age 2-3 reduces 40% of one standard deviation of Behavior Problems Index (BPI) at age 6-7. In comparison, the effects of spanking at age 4-5 are small and ambiguous in sign. These results disagree with the prevailing findings in the psychology/education literature where the empirical findings are not backed by a proper causal framework. We hope our approach to be applied to other dynamic causal relations. This will be taking one step further from the simple Granger causality analysis toward the full causal analysis allowing for feedbacks from interim responses.

APPENDIX

Proof for (3.2)

NUC (i): for '$\Longleftarrow$', observe

$$\begin{aligned} & E(y_1^j|d_1=j,y_0,x_1) = E(y_1|d_1=j,y_0,x_1) \quad \text{because } y_1 = y_1^j \text{ given } d_1 = j \\ = \; & g(j,y_0,x_1) \quad \text{from } E(y_1|d_1,y_0,x_1) = g(d_1,y_0,x_1) \\ = \; & E(y_1^j|y_0,x_1) \quad \text{from } E(y_1^j|y_0,x_1) = g(j,y_0,x_1) \text{ in (3.1).} \end{aligned}$$

For '$\Longrightarrow$', observe, with $1[d_1 \le j]$ denoting the distribution for j degenerate at d_1,

$$\begin{aligned} & E(y_1|d_1,y_0,x_1) = \int E(y_1|j,y_0,x_1)\partial 1[d_1 \le j] = \int E(y_1^j|j,y_0,x_1)\partial 1[d_1 \le j] \\ = \; & \int E(y_1^j|y_0,x_1)\partial 1[d_1 \le j] = \int g(j,y_0,x_1)\partial 1[d_1 \le j] = g(d_1,y_0,x_1) \end{aligned}$$

where ∂ is used instead of d for integration to prevent confusion.

NUC (ii): for '$\Longleftarrow$', observe

$$\begin{aligned}
&E(y_2^{jk}|d_1=j,d_2=k,y_1,x_2)=E(y_2|d_1=j,d_2=k,y_1,x_2)\\
=\;&h(j,k,y_1,x_2)\quad\text{from } E(y_2|d_1,d_2,y_1,x_2)=h(d_1,d_2,y_1,x_2)\\
=\;&E(y_2^{jk}|y_1,x_2)\quad\text{from } E(y_2^{jk}|y_1,x_2)=h(j,k,y_1,x_2).
\end{aligned}$$

For '$\Longrightarrow$', observe, with $1[d_1\le j,d_2\le k]$ denoting the distribution for (j,k) degenerate at (d_1,d_2),

$$\begin{aligned}
&E(y_2|d_1,d_2,y_1,x_2)=\int E(y_2|j,k,y_1,x_2)\partial 1[d_1\le j,d_2\le k]\\
=\;&\int E(y_2^{jk}|j,k,y_1,x_2)\partial 1[d_1\le j,d_2\le k]=\int E(y_2^{jk}|y_1,x_2)\partial 1[d_1\le j,d_2\le k]\\
=\;&\int h(j,k,y_1,x_2)\partial 1[d_1\le j,d_2\le k]=h(d_1,d_2,y_1,x_2).
\end{aligned}$$

Proof for G estimation Identifying Total Effect in Two Periods

Use $y_2^{jk}=\beta_1+\beta_{d1}j+\beta_{d2}k+\beta_y y_1+\beta_{x2}'x_2+v_2$ to get

$$\begin{aligned}
&E(y_2^{jk}|d_1=j,d_2=k,\ y_1,y_0,X_2)=E(y_2^{jk}|d_1=j,\ y_1,y_0,X_2)\quad\text{owing to NUC (b)}\\
=\;&\beta_1+\beta_{d1}j+\beta_{d2}k+\beta_y y_1+\beta_{x2}'x_2+E(v_2|d_1=j,\ y_1,y_0,X_2).
\end{aligned}$$

Integrate out y_1 conditional on $(d_1=j,y_0,X_2)$ following G estimation to get

$$E(y_2^{jk}|d_1=j,y_0,X_2)=\beta_1+\beta_{d1}j+\beta_{d2}k+\beta_y E(y_1^j|d_1=j,y_0,X_2)+\beta_{x2}'x_2+E(v_2|d_1=j,y_0,X_2).$$

Due to NUC (a), taking $E(\cdot|d_1=j,y_0,X_2)$ is the same as taking $E(\cdot|y_0,X_2)$. Thus the last display equals

$$E(y_2^{jk}|y_0,X_2)=\beta_1+\beta_{d1}j+\beta_{d2}k+\beta_y E(y_1^j|y_0,X_2)+\beta_{x2}'x_2+E(v_2|y_0,X_2).$$

From the y_1^j equation, we get

$$E(y_1^j|y_0,X_2)=\gamma_1+\gamma_d j+\gamma_y y_0+\gamma_{x1}'x_1+E(v_1|y_0,X_2).$$

Substitute this into the preceding display to get

$$\begin{aligned}
&\beta_1+\beta_{d1}j+\beta_{d2}k+\beta_y\{\gamma_1+\gamma_d j+\gamma_y y_0+\gamma_{x1}'x_1+E(v_1|y_0,X_2)\}+\beta_{x2}'x_2+E(v_2|y_0,X_2)\\
=\;&(\beta_1+\beta_y\gamma_1)+(\beta_{d1}+\beta_y\gamma_d)j+\beta_{d2}k+\beta_y\gamma_y y_0+\beta_y\gamma_{x1}'x_1+\beta_{x2}'x_2+E(\beta_y v_1+v_2|y_0,X_2).
\end{aligned}$$

From this, as desired,

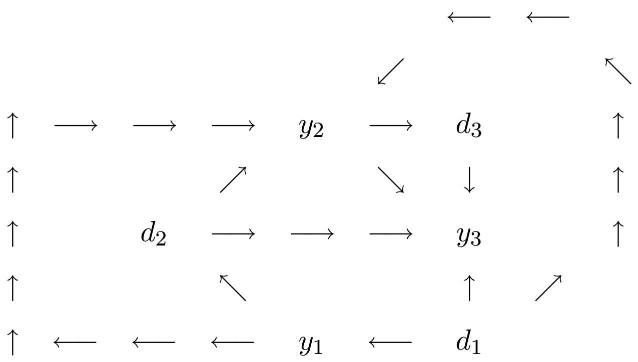
$$E(y_2^{jk}|y_0, X_2) - E(y_2^{00}|y_0, X_2) = (\beta_{d1} + \beta_y \gamma_d)j + \beta_{d2}k.$$

Extension of Structural IVE to Three Periods/Treatments

The treatment profile is $d = (d_1, d_2, d_3)'$ and the observation sequence is

$$(x_0, y_0),\ (d_1, \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}),\ (d_2, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}),\ (d_3, \begin{pmatrix} x_3 \\ y_3 \end{pmatrix}).$$

The desired effect is $E(y_3^{jkl} - y_3^{000})$ and the direct and indirect effects are:



The linear (contemporaneous covariate) models are

$$\begin{aligned} y_1^j &= \gamma_{11} + \gamma_{1d1} j + \gamma_{1y} y_0 + \gamma_{1x}' x_1 + v_1, \\ y_2^{jk} &= \gamma_{21} + \gamma_{2d1} j + \gamma_{2d2} k + \gamma_{2y} y_1^j + \gamma_{2x}' x_2 + v_2, \\ y_3^{jkl} &= \beta_1 + \beta_{d1} j + \beta_{d2} k + \beta_{d3} l + \beta_y y_2^{jk} + \beta_{x3}' x_3 + v_3. \end{aligned}$$

The notations differ somewhat from the two period case, for y_3 is the final response. The y_3^{jkl} reduced form (RF) with both y_{i2}^{jk} and y_{i1}^j removed is

$$\begin{aligned} y_3^{jkl} = {} & \{\beta_1 + \beta_y(\gamma_{21} + \gamma_{2y}\gamma_{11})\} + \{\beta_{d1} + \beta_y(\gamma_{2d1} + \gamma_{2y}\gamma_{1d1})\}j + (\beta_{d2} + \beta_y \gamma_{2d2})k + \beta_{d3} l \\ & + \beta_y \gamma_{2y}\gamma_{1y} y_0 + \beta_y \gamma_{2y} \gamma_{1x}' x_1 + \beta_y \gamma_{2x}' x_2 + \beta_{x3}' x_3 + (\beta_y \gamma_{2y} v_1 + \beta_y v_2 + v_3). \end{aligned}$$

This shows five effects to be identified:

$$\begin{aligned} \text{direct and indirect (through } y_1, y_2\text{) effects of } d_1 &: \ \beta_{d1},\ \ \beta_y(\gamma_{2d1} + \gamma_{2y}\gamma_{1d1}) \\ \text{direct and indirect (through } y_2\text{) effects of } d_2 &: \ \beta_{d2},\ \ \beta_y \gamma_{2d2} \\ \text{direct effect of } d_3 &: \ \beta_{d3}. \end{aligned}$$

As in the two-period case, all three period equations may be estimated, or alternatively under equal contemporaneous effect assumption $\gamma_{1d1} = \gamma_{2d2}$, only the y_2 and y_3 equations can be estimated. Going further, under

$$\beta_{d3} = \gamma_{1d1} = \gamma_{2d2}, \quad \gamma_{2y} = \beta_y, \quad \gamma_{2d1} = \beta_{d2},$$

only the y_3 equation has to be estimated.

Covariate Choice and Instruments

Choosing a model, classifying the covariates as exogenous or endogenous, and then finding instruments for the endogenous regressors are a complicated endeavor in observational data where experiments may be infeasible as in spanking. At best of times, these processes are "controversial but convincing". Here we provide the details of our model building. Given the subtle nature of spanking—even showing (no) emotion while spanking can result in a world of difference—making our procedure as scientific as the reader would like might be difficult, in which case the reader may take our empirical analysis just as an illustration of the proposed method.

In our data, there is no known time order between x_t and y_t; with temporal aggregation, x_t and y_t can be simultaneously related. This raises the issue of which covariates to include in the y_1 and y_2 equations. A component w_1 of x_1 may be affected by y_1 or d_1. In such a case, should w_1 be still included in x_1? We examine this issue here, assuming that w_1 *affects* y_1; if w_1 affects y_2 but not y_1, w_1 should be put into x_2; if w_1 does affect neither y_1 nor y_2, then w_1 can be simply ignored.

First, suppose that w_1 is affected by d_1, but not by y_1. If w_1 is included in x_1, then the indirect effect $d_1 \rightarrow w_1 \rightarrow y_1$ is missed because w_1 is controlled; if interested only in the direct effect, however, then including w_1 in x_1 is all right. If we choose not to include w_1 in x_1 to avoid this problem, then we may incur another problem. For instance, there may be an unobserved common factor affecting both d_1 and w_1; with w_1 becoming part of the error term, this makes d_1 endogenous. Thus it seems safer to control w_1. For example, suppose that w_1 is 'reading (books) to children'. A parent may do this because of a guilty feeling after spanking (hence d_1 affects w_1), which then influences y_1. Controlling reading-to-children entails missing this indirect effect. But not controlling it may entail a confounding as just mentioned. Plus, when one talks about 'effects of spanking', it is more likely than not that the effects refer to those when the other home input variables such as reading-to-children are controlled. Hence, in our empirical analysis, we include variables such as reading-to-children in the y_1 equation, assuming either that there is no w_1 affected by d_1, or that, if there is such a w_1, then we are not interested in the indirect effect. In G4 of Table 0, we classify all variables except the last 8 'punishment' variables as this type of variables (analogously, for y_2, all variables in G3 except the last 7 punishment/reward variables are classified as w_2).

Second, suppose that w_1 is affected by y_1. In this case, w_1 gets simultaneously related to y_1 and becomes an endogenous regressor in the y_1 equation. For instance, various disciplinary measures (e.g., grounding or taking away allowances) can be simultaneously related to y_1 (due to the temporal aggregation). Consider thus a bivariate simultaneous equation model where (w_1, y_1) becomes the bivariate response variables. In the y_1 structural form (SF), the coefficient of d_1 shows only the direct effects (as if an intervention on d_1 is accompanied by an intervention on w_1). In contrast, in the y_1 reduced form (RF) with w_1 substituted out, the coefficient of d_1 shows the total effect. For instance, suppose

$$y_1 = \alpha_w w_1 + \alpha_d d_1 + u, \quad w_1 = \beta y_1 + \varepsilon \implies y_1 = \frac{\alpha_d}{1 - \alpha_w \beta} d_1 + \frac{u + \alpha_w \varepsilon}{1 - \alpha_w \beta} \quad \text{where } |\alpha_w \beta| < 1.$$

In words, an initial change in d_1 causes a change in y_1 of magnitude α_d, but the change in y_1 leads to a change in w_1 of magnitude β, which in turn changes y_1 and so on. The y_1 RF includes the full effect of d_1 coming from the repeated exchanges between y_1 and w_1 that were triggered by the initial d_1 change ('$|\alpha_w \beta| < 1$' is for the repeated exchanges to "converge").

In our empirical analysis, we try both including and excluding the simultaneously related variables. Including those endogenous variables and estimating the y_1 SF with IVE means that the estimated effect of d_1 is only the direct effect without any other disciplinary/rewarding measures taken to substitute for or complement d_1. Excluding those variables means that we are estimating the y_1 RF where the total effect of d_1 gets estimated.

Specifically, let m_2 denote the variables in G3 of Table 0 other than the last 7 punishment/reward variables which are denoted as s_2. Also let m_1 denote the variables in G4 other than the last 8 punishment variables (not including HOME) which are denoted as s_1. Then two models are used:

$$(i) \;:\; y_1 = \text{linear fn. } m_1, s_1, G1, G2 + u_1 \quad \text{and} \quad y_2 = \text{linear fn. } m_2, s_2, G1, G2 + u_2$$

$$(ii) \;:\; y_1 = \text{linear fn. } m_1, \quad G1, G2 + u_1 \quad \text{and} \quad y_2 = \text{linear fn. } m_2, \quad G1, G2 + u_2$$

For (i), IVE is necessary for the endogenous s_1 and s_2; we use G5 for s_1, and $(G5', m_1', s_1')'$ for s_2. For (ii), we use LSE.

The same issue of covariate choice arises for the y_2 equation. In principle, one just has to follow the same model as used for the y_1 equation but augmented by d_2 now, although this could not be done exactly with our data as different sets of variables were available for the y_1 and y_2 equations.

REFERENCES

Angrist, J.D. and G.W. Imbens, 1995, Two-stage least squares estimation of average causal effects in models with variable treatment intensity, Journal of the American Statistical Association 90, 431-442.

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, 444-455.

Baumrind, D., R.E. Larzelere, and P.A. Cowan, 2002, Ordinary physical punishment: Is it harmful? Comment on Gershoff (2002), Psychological Bulletin 128, 580-589.

Gershoff, E., 2002, Corporal punishment by parents and associated child behaviors and experiences: a meta-analytic and theoretical review, Psychological Bulletin 128, 539–579.

Granger, C.W.J., 1969, Investigating causal relations by econometric models and cross-spectral methods, Econometrica 37, 424-438.

Granger, C.W.J., 1980, Testing for causality: a personal viewpoint, Journal of Economic Dynamics and Control 2, 329-352.

Holland, P.W., 1986, Statistics and causal inference, Journal of the American Statistical Association 81, 945-960.

Imbens, G.W. and D.B. Rubin, 1997, Bayesian inference for causal effects in randomized experiments with noncompliance, Annals of Statistics 25, 305-327.

Lee, M.J., 2000, Median treatment effect in randomized trials, Journal of the Royal Statistical Society (Series B) 62, 595-604.

Lee, M.J., 2002, Panel data econometrics: methods-of-moments and limited dependent variables, Academic Press

Lee, M.J., 2005, Micro-econometrics for policy, program, and treatment effects, Oxford University Press.

Pearl, J., 2000, Causality, Cambridge University Press.

Robins J. M., 1986, A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect, Mathematical Modelling 7, 1393-1512.

Robins, J.M., 1992, Estimation of the time-dependent accelerated failure time model in the presence of confounding factors, Biometrika 79, 321-334.

Robins, J.M., 1998, Structural nested failure time models, in Survival Analysis, Vol. 6, Encyclopedia of Biostatistics, edited by P. Armitage and T. Colton, Wiley.

Robins, J.M., 1999, Marginal structural models versus structural nested models as tools for causal inference, in Statistical models in epidemiology: the environment and clinical trials, edited by M.E. Halloran and D. Berry, Springer, 95-134.

Robins, J.M., S. Greenland, and F.C. Hu, 1999, Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome, Journal of the American Statistical Association 94, 687-700.

Rosenbaum, P., 2002, Observational studies, 2nd ed., Springer.

Rubin, D.B., 1974, Estimating causal effects of treatments in randomized and nonrandomized studies, Journal of Educational Psychology 66, 688-701.

Rubin, D.B., 2005, Causal inference using potential outcomes: design, modeling, decisions, Journal of the American Statistical Association 100, 322-331.

Witteman, J.C.M., R.B. D'Agostino, T. Stijnen, W.B. Kannel, J.C. Cobb, M.A.J de Ridder, A. Hofman, and J.M. Robins, 1998, G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham heart study, American Journal of Epidemiology 148, 390-401.