*The Korean Economic Review*
Volume 36, Number 1, Winter 2020, 135-173.

# Review on Difference in Differences*

Myoung-jae Lee** · Yasuyuki Sawada***

*Difference in differences (DD) is one of the most popular approaches in economics and other disciplines of social sciences. This paper provides a review on the basics and recent advances in DD from a personal perspective. Details on DD identification and estimation using panel data and repeated cross-sections are provided for various DD cases such as constant/time-varying effect or constant/time-varying treatment timing. Following these basics on DD, topics such as 'DD in reverse', fuzzy DD, synthetic control, and triple and generalized differences are examined. Many empirical examples in various areas of economics are provided for illustration.*

JEL Classification: C21, C23, H00, I00, J08, O10
Keywords: Difference in Differences (DD), DD in Reverse, Fuzzy DD, Synthetic Control, Triple DD, Generalized DD

## I. Introduction

An article (November 26, 2016) in *The Economist*, entitled "Economists are prone to Fads, ...", analyzed the most frequently used techniques in economics by a machine learning technique. The analysis was based on key words in the abstracts of NBER working papers, and the most popular methods turned out to be difference in differences (DD), followed by regression discontinuity (RD), laboratory experiment, dynamic stochastic general equilibrium, randomized control trial, and machine-learning/big-data. DD has been at top since 2012 and its

*Received: Oct. 1, 2018. Revised: Jan. 16, 2019. Accepted: March 8, 2019.*

* The authors are grateful to two anonymous reviewers for their helpful comments. Also the feedback from the DD lecture audiences at Asian Development Bank, Australian National University, Korea Development Institute, Korea Institute of Public Finance, Seoul National University, and University of Luxembourg led to greatly improving the paper.

** First and Corresponding Author, Professor, Dept. Economics, Korea University, Seoul 02841, South Korea, Phone/Fax: 82-2-3290-2229, Email: myoungjae@korea.ac.kr

*** Second Author, Chief Economist, Asian Development Bank, and Professor, Faculty of Economics, University of Tokyo, Email: ysawada@adb.org

popularity has been increasing ever since, unlike some other methods. There are various references for DD: see Angrist and Krueger (1999), Heckman et al. (1999), Lee (2005), Athey and Imbens (2006), Angrist and Pischke (2009), Lee (2016a), and references therein.

Despite the popularity of DD, Besley and Case (2000) warned of endogeneity problems in a policy/treatment, and suggested to explore the policy equation to find plausible instruments in the political process determining the policy, such as the number of women or minorities in congress who might be keener on family/health-related policies. Also, Bertrand et al. (2004) illustrated DD inference problems involving 'clustering/grouping' that observations are related to one another by sharing the individual index $i$ (i.e., belonging to the same individual), the time index $t$, or something else such as age or residential area. A treatment varying only at an aggregate level, not at the individual level, raises another inferential problem. The policy endogeneity issue is not dealt with in this paper, however, as it is not unique to DD. Also, we do not address the DD inference problems to keep this paper not too long; interested readers may refer to Lee (2016a), Brewer et al. (2018), and references therein, where the main message seems to be "use at least panel generalized least squares estimator with a clustered variance estimator to account for serial correlations and others".

In the rest of this introductory section, first, we present the basics of DD and two illustrative examples. Second, our notation to be used throughout this paper is introduced. Third, we lay out how the simple 'before-after' is generalized/related to DD and then to triple differences, hoping to give the reader a "big picture" for this paper.

### 1.1. Basics and Two Examples

In the basic setup of DD, there are two groups based on a time-constant treatment-qualification/eligibility dummy $Q_i$: the $Q_i = 1$ group, i.e., 'treatment group (T group)', and the $Q_i = 0$ 'control group (C group)'. In DD, the T group gets treated at a time point, say $\tau$, (and onwards), but the C group never. Bear in mind that the T group is not treated before $\tau$, despite that it is called a T group.

A number of variations of the basic DD framework arise. First, the treatment timing, say $\tau_i$, can vary across individuals $i = 1,\ldots,N$. Second, $Q_i$ can be time-constant (e.g., gender and race), or time-varying as in $Q_{it} = 1$ for income/wealth below some threshold or the number of children above zero. Third, the C group may be always treated instead of always untreated ('DD in reverse'); e.g., the C group is a village already with a bridge (a treatment) across a river, and the T group has a bridge constructed later than the C group. Despite these variations, we will stick to the basic setup with two periods, $\tau$ and $Q_i$ for a while unless otherwise mentioned, and address the extensions/generalization later.

DD requires at least two waves of observations. Hence DD can be implemented with either repeated cross-sections (RCS) or panel data. As well known, panel data can deal with various treatment endogeneity issues better than RCS can; also DD Identification (ID) can be seen more easily with panel data. However, estimation with panel data is more involved, because RCS is basically handled as cross-section data with each individual observed only once over time. For RCS, ordinary least squares estimator (OLS) or instrumental variable estimator (IVE) can be applied. That is, there are pros and cons in using panel data vs. RCS for DD.

Consider a three-strike law example. The 'three-strike law' in California (CA) was enacted to lower crime rates: if convicted 3 times, the person is jailed for life; see Helland and Tabarrok (2007) for more on three strike laws. The effect may be seen by comparing the crime rates of CA, say, 1 year before and after the law, which is a before-and-after (BA). In the study period, however, many other things can change. For instance, the CA economy may improve to lower the crime rate. One way to remove the undesired economy/time effect is finding a control state, say Washington (WA) state, that did not have the treatment but experienced the same change in the economic/time conditions. The crime rate BA of CA contains both the economy and three-strike-law effects, whereas the crime rate BA of WA contains only the economy effect. The difference of the two BA's (i.e., DD) yields the desired treatment effect. Since WA is selected for its economic/time conditions being similar to those of CA, *DD essentially combines BA with matching*.

To give specific numbers, consider another example. Immigration of cheap labor (treatment) is blamed for minority unemployment (response/outcome). Miami experienced an influx of cheap labor from Cuba through the "boatlift" incidence 1979-1981. During the period, the Miami unemployment rate has increased by 1.3%. Card (1990) explained the boatlift incidence—the change in the U.S. policy handling Cuban refugees, Castro releasing criminals and mental patients, ...—and used 4 control cities: Atlanta, Houston, LA and Tampa. The four cities are chosen because they are thought to share the same time effect with Miami; if not, a weighted average of the four cities may make a better control group, which is called a 'synthetic control'. Using the Current Population Survey, one result from Card (1990) in Table 'DD for Immigration Effect on Unemployment' shows that the control states experienced even higher unemployment to result in an insignificant unemployment-decreasing effect.

DD for Immigration Effect on Unemployment

| | 1979 | 1981 | 1981-1979 BA (SE) |
|---|---|---|---|
| Miami | 8.3 | 9.6 | 9.6−8.3 = 1.3 (2.5) |
| Control Group Average | 10.3 | 12.6 | 12.6−10.3 = 2.3 (1.2) |
| DD Treatment Effect | | | 1.3−2.3 = −1.0 (2.8) |

### 1.2. Notation

Introducing notation, consider two periods $t=2,3$ with a time-constant qualification $Q_i$, where the treatment is applied only to the $Q_i=1$ group at $t=3$. There are reasons for using $t=2,3$, instead of $t=1,2$ or $t=0,1$: (i) to avoid confusion with $Q_i=0,1$, (ii) to allow the period-1 response as a regressor in period 2, (iii) and to consider later the difference between DD over periods 2-3 and DD over periods 1-2.

In most DD cases, the treatment $D_{it}$ is the interaction of $Q_i$ and $1[t=3]$ where $1[A]=1$ if $A$ holds and 0 otherwise:

$$D_{it} \equiv Q_i 1[t=3].$$

There are cases with $D_{it} \neq Q_i 1[t=3]$, because one does not necessarily have to be treated even if $Q_i 1[t=3]=1$, or may be treated when $Q_i 1[t=3] \neq 1$. Here, $Q_i 1[t=3]=1$ is just an eligibility to get treated, and we call this 'fuzzy DD'.

Let $Y_{it}^1$ it be the 'potential' treated response of individual $i$ at time $t$, and $Y_{it}^0$ the potential untreated response. The observed response is

$$Y_{it}=(1-D_{it})Y_{it}^0+D_{it}Y_{it}^1 \Rightarrow Y_{i2}=Y_{i2}^0, \;\; Y_{i3}=(1-Q_i)Y_{i3}^0+Q_iY_{i3}^1.$$

Define

$$\Delta Y_{it} \equiv Y_{it}-Y_{i,t-1}, \;\; \Delta Y_{it}^0 \equiv Y_{it}^0-Y_{i,t-1}^0, \;\; \Delta Y_{it}^1 \equiv Y_{it}^1-Y_{i,t-1}^1.$$

Let the time-constant and time-varying covariates be $C_i$ and $X_{it}$,

$$W_{it} \equiv (C_i', X_{it}')' \;\text{ and }\; W_{i,t-1}^t \equiv (C_i', X_{i,t-1}', X_{it}')'.$$

For simplicity, covariates are often omitted, which can be confusing at times, as the omitted covariates may refer to different periods. Also, the subscript $i$ indexing individuals is often omitted as in writing $Q_i$ as $Q$. Furthermore, both subscripts $i$ and $t$ are omitted occasionally as in writing $D_{it}$ and $Y_{it}$ just as $D$ and $Y$.

### 1.3. From BA to DD and to TD

In BA which is the most basic method in causal analysis, the same subjects are compared before and after a treatment $D$; e.g., BA photos of persons with a plastic surgery $D$. Formally, BA for the $Q=1$ group is

$$E(\Delta Y_3 \mid Q=1) = E(Y_3^1 \mid Q=1) - E(Y_2^0 \mid Q=1) \tag{1.1}$$
$$= E(Y_3^1 \mid Q=1) - E(Y_3^0 \mid Q=1) + \{E(Y_3^0 \mid Q=1) - E(Y_2^0 \mid Q=1)\} \tag{1.2}$$

subtracting and adding $E(Y_3^0 \mid Q=1)$ after the first term. $E(Y_3^0 \mid Q=1)$ is a counter-factual because $Y_3^0$ is never realized for $Q=1$.

The BA ID condition is that the terms in $\{\cdot\}$ are zero:

$$E(\Delta Y_3^0 \mid Q=1)\{= E(Y_3^0 \mid Q=1) - E(Y_2^0 \mid Q=1)\} = 0 \qquad (\mathrm{ID}_{BA})$$

which makes BA in (1.2) '*the effect on the treated* $Q=1$ *at the post-treatment period*'

$$E(Y_3^1 - Y_3^0 \mid Q=1). \tag{1.3}$$

Now suppose that the treatment takes a long time to manifest itself, during which other variables (observed $X$ and unobserved $\varepsilon$) can change. Then the change in $Y$ may be due to changes in $X$ or $\varepsilon$, not necessarily due to $D$; call the changes due to $X$ or $\varepsilon$ the "time effect/trend". Then we use DD which is the BA of the T group minus the BA of the C group:

$$E(\Delta Y_3 \mid Q=1) - E(\Delta Y_3 \mid Q=0); \tag{1.4}$$

the BA in (1.1) is the first half of the DD, which experiences both the time and treatment effects. The role of the second BA in DD is to remove the time effect from the first BA, using the C group that experiences the same time effect, but not the treatment itself. DD thus yields the desired treatment effect.

When we select a $Q=0$ group for DD, we should select a group that shares the same time effect with the $Q=1$ group so that the time effect lurking in the BA of the T group can be removed by the BA of the C group. This aspect of selecting a $Q=0$ group that is similar to the $Q=1$ group in time effect can be called 'matching', as was already mentioned. We can check if the BA of the C group is zero (i.e., $E(\Delta Y_3 \mid Q=0)=0$), in which case DD reduces to the BA of the $Q=1$ group.

As will be seen in detail later, the DD ID condition is

$$E(\Delta Y_3^0 \mid Q=1) = E(\Delta Y_3^0 \mid Q=0) \qquad (\mathrm{ID}_{DD})$$

which generalizes the above $\mathrm{ID}_{BA}$, because $\mathrm{ID}_{DD}$ allows $E(\Delta Y_3^0 \mid Q=1)$ not to be zero as long as it is the same as $E(\Delta Y_3^0 \mid Q=0)$. Strictly speaking, however, this is not a generalization, because $\mathrm{ID}_{BA}$ does not involve the $Q=0$ group while

$\text{ID}_{DD}$ does. DD will be shown to identify $E(Y_3^1 - Y_3^0 \mid Q=1)$ in (1.3) as BA does.

Although $\text{ID}_{DD}$ is the main DD ID condition to be invoked throughout this paper, a different DD ID condition also appears:

$$E(\Delta Y_3^0) = E(\Delta Y_2^0) \quad \text{or} \quad E(\Delta Y_3^0 \mid Q=1) = E(\Delta Y_2^0 \mid Q=1). \qquad (\text{ID}_{DDt})$$

$\text{ID}_{DDt}$ requires equality along the time dimension ($t=3$ vs. $t=2$), whereas $\text{ID}_{DD}$ requires equality across the cross-sectional group dimension ($Q=1$ vs. $Q=0$).

Going further, we can think of 'triple difference (TD)' which is a difference of two DD's to allow $E(\Delta Y_3^0 \mid Q=1) \neq E(\Delta Y_3^0 \mid Q=0)$. Given another group dummy $G$, the ID condition for TD will be shown to be

$$\begin{aligned} &E(\Delta Y_3^0 \mid G=1, Q=1) - E(\Delta Y_3^0 \mid G=1, Q=0) \\ &= E(\Delta Y_3^0 \mid G=0, Q=1) - E(\Delta Y_3^0 \mid G=0, Q=0). \end{aligned} \qquad (\text{ID}_{TD})$$

This generalizes $\text{ID}_{DD}$ because $\text{ID}_{TD}$ allows $E(\Delta Y_3^0 \mid Q=1) \neq E(\Delta Y_3^0 \mid Q=0)$ with $G=0,1$ groups. Again, strictly speaking, $\text{ID}_{TD}$ is not a generalization of $\text{ID}_{DD}$, because $\text{ID}_{DD}$ does not involve $G$ whereas $\text{ID}_{TD}$ does.

There is a TD which adds the extra DD along the time dimension, not along the cross-sectional group ($G$) dimension, and its ID condition is

$$E(\Delta Y_3^0 \mid Q=1) - E(\Delta Y_3^0 \mid Q=0) = E(\Delta Y_2^0 \mid Q=1) - E(\Delta Y_2^0 \mid Q=0). \qquad (\text{ID}_{TDt})$$

The relationship between $\text{ID}_{TD}$ and $\text{ID}_{TDt}$ is analogous to that between $\text{ID}_{DD}$ and $\text{ID}_{DDt}$. Although we introduced various ID conditions for BA, DD and TD using panel data (i.e., $\Delta Y$ is used), analogous ID conditions exist for RCS.

# II. DD with Panel Data

## 2.1. Identification with Panel Data

Differently from (1.2) for BA where we subtracted and added the counterfactual $E(Y_3^0 \mid Q=1)$, subtract and add the counterfactual $E(\Delta Y_3^0 \mid Q=1)$ after the first term of

$$DD_{23} \equiv E(\Delta Y_3 \mid Q=1) - E(\Delta Y_3 \mid Q=0)$$

to obtain, due to $E(\Delta Y_3 \mid Q=1) = E(Y_3^1 - Y_2^0 \mid Q=1)$,

$$DD_{23} = E(Y_3^1 - Y_2^0 \mid Q=1) - E(\Delta Y_3^0 \mid Q=1) + \{E(\Delta Y_3^0 \mid Q=1) - E(\Delta Y_3^0 \mid Q=1)\} .$$

The terms in $\{\cdot\}$ drop out under $\text{ID}_{DD}$ to make $DD_{23}$ equal to (1.3)—the effect on the treated at the post-treatment period—which BA also becomes under $\text{ID}_{BA}$:

$$DD_{23} = E(Y_3^1 - Y_2^0 \mid Q=1) - E(Y_3^0 - Y_2^0 \mid Q=1) = E(Y_3^1 - Y_3^0 \mid Q=1) .$$

This is natural, because the effect is seen only for those whose $D_{it}$ changes (i.e., $Q=1$) and only at the time when $D_{it}$ changes (i.e., at $t=3$).

$\text{ID}_{DD}$, often called the 'same time effect' or 'parallel trend' assumption, is that $Q$ is as good as randomized for $\Delta Y_3^0$. If multiple pre-treatment periods are available, then the $Y$ paths of the T and C groups in the pre-treatment periods can be presented to graphically demonstrate $\text{ID}_{DD}$: the paths should be "parallel". This parallelism should not be taken literally, because what is needed is the two $Y$ paths moving together, possibly only with the same vertical difference over time. When the parallel paths do not hold, if there are multiple control groups whose convex combination can give a parallel path, then the combination may be used as a single control group, called 'synthetic control' (see Abadie et al., 2015 and references therein).

To illustrate parallel paths, we simulated some data: with $N=1000$, $T=45$ (periods in total) and $\tau=23$ (treatment starting halfway), let

$$Y_{it} = (1+0.5t-0.01t^2) + Q_i + 2X_{it} + 3Q_i 1[\tau \le t] + U_{it},\ U_{it} \sim N(0,1),$$
binary $Q_i$ with equal probability, $X_{it} \mid (Q_i=0) \sim U[0,1]$,
$X_{it} \mid (Q_i=1) \sim U[0,2]$

where the time effect is $1+0.5t-0.01t^2$, and the $X_{it}$'s distribution differs between the $Q=0$ and $Q=1$ groups.

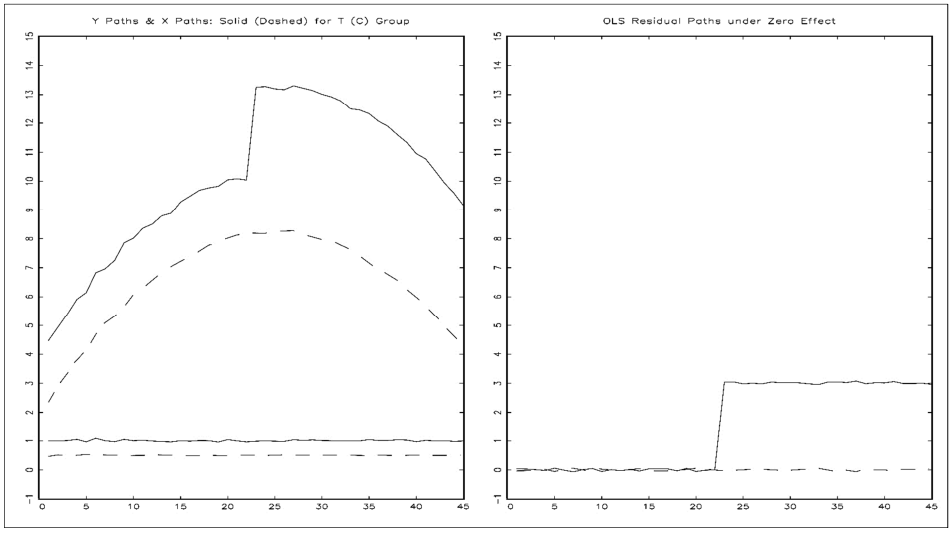
The left panel of Figure 'Parallel Paths of $Y$, $X$ and OLS Residual Mean' plots quadratic $E(Y_{it} \mid Q_i)$ for each t with the upper two lines, and flat $E(X_{it} \mid Q_i)$ with the bottom two lines. The pre-treatment paths are parallel with the vertical difference 2, which is the sum of the $Q$ slope 1 and $2 \times \{E(X_{it} \mid Q_i=1) - E(X_{it} \mid Q_i=0)\} = 1$, and the treatment effect 3 that is the slope of $Q_i 1[\tau \le t]$ is clearly visible at $t=23$ in the figure not controlling $X$ despite the $X$ distribution difference between the two groups. For the right panel, we do the OLS of $Y_{it}$ on

$$1, 1[t=2], \ldots, 1[t=T],\ Q_i,\ X_{it},\ Q_i 1[\tau \le t],$$

but then plot the OLS residual mean computed setting $Q_i 1[\tau \le t] = 0$ for the

$Q=1$ group so that the plot includes the jump magnitude 3 at $t=23$ (solid line); the OLS residual mean for the $Q=0$ group is also plotted (dashed line).

**[Figure 1]** Parellel Paths of $Y$, $X$ and OLS Residual Mean




Although we omitted $W_2^3$ in the conditioning set so far, we may want to make its presence explicit by writing $\mathrm{ID}_{DD}$ as

$$E(\Delta Y_3^0 \mid W_2^3, Q=1) = E(\Delta Y_3^0 \mid W_2^3, Q=0): \tag{2.1}$$

$Q$ is as good as randomized for $\Delta Y_3^0$ given $W_2^3$. That is, the part other than $W_2^3$ in $\Delta Y_3^0$ (i.e., the error term in $\Delta Y_3^0$) is balanced across the two groups. The error term in $\Delta Y_3^0$ is allowed to be related to $W_2^3$, as long as the relationship is the same across the two groups. With $W_2^3$ in the conditioning set, $DD_{23}$ becomes the $W_2^3$-conditional effect $E(Y_3^1 - Y_3^0 \mid W_2^3, Q=1)$.

If we desire a marginal version free of $W_2^3$, then $W_2^3$ should be integrated out. Using the distribution $F_{W_2^3|Q=1}$ of $W_2^3 \mid Q=1$, we obtain

$$E(Y_3^1 - Y_3^0 \mid Q=1) = \int E(Y_3^1 - Y_3^0 \mid W_2^3 = w, Q=1) dF_{W_2^3|Q=1}(w).$$

We may use another distribution in this integration, such as $F_{W_2^3|Q=0}$ or $F_{W_2^3}$. But then, the resulting integral is something else, not $E(Y_3^1 - Y_3^0 \mid Q=1)$.

An alternative to $\mathrm{ID}_{DD}$ is "zero effect of $Q$ on $Y_3^1 - Y_2^0$":

$$E(Y_3^1 - Y_2^0 \mid Q=1) = E(Y_3^1 - Y_2^0 \mid Q=0). \tag{$\mathrm{ID}'_{DD}$}$$

This $\mathrm{ID}'_{DD}$ involving both $Y^1$ and $Y^0$ is, however, deemed to be less plausible than $\mathrm{ID}_{DD}$ involving only $Y^0$. Under $\mathrm{ID}'_{DD}$, $DD_{23}$ becomes the 'effect on the untreated $Q=0$ at $t=3$':

$$E(Y_3^1 - Y_3^0 \mid Q=0).$$

If both $\mathrm{ID}_{DD}$ and $\mathrm{ID}'_{DD}$ hold, then $DD_{23}$ becomes the 'effect on the population at $t=3$' $E(Y_3^1 - Y_3^0)$, because $DD_{23}$ equals both effects on the right-hand side of

$$E(Y_3^1 - Y_3^0) = E(Y_3^1 - Y_3^0 \mid Q=0)P(Q=0) + E(Y_3^1 - Y_3^0 \mid Q=1)P(Q=1).$$

If we make the presence of $W_2^3$ explicit, $DD_{23}$ equals $E(Y_3^1 - Y_3^0 \mid W_2^3)$, and integrating out $W_2^3$ with $F_{W_2^3}$ gives $E(Y_3^1 - Y_3^0)$. For simplicity, we will not further mention this aspect of conditioning on $W_2^3$ first and then integrating it out later. Further discussion on DD identification can be found in Lee and Kang (2006) and Lee (2016a).

## 2.2. Estimation with Panel Linear Models

Although we discussed ID using conditional means such as $E(Y \mid Q=1)$ omitting $W_2^3$, if we make the presence of $W_2^3$ explicit, then estimating conditional means such as $E(Y \mid W_2^3, Q=1)$ to implement DD is difficult unless the functional form is specified. For a two-wave panel data, for simplicity, we may use a panel linear model:

$$Y_{it} = \beta_t + \beta_q Q_i + \beta_d Q_i 1[t=3] + \beta_w' W_{it} + \delta_i + U_{it}$$

where $\beta_t$ is a time-varying intercept, $\beta_q$ is the group $(Q=1)$ effect, $\beta_d$ is the treatment effect, $\beta_w$ is the slope of $W_{it}$, $\delta_i$ is an unit-specific error, and $U_{it}$ is an unit- and time-varying error.

The simplest estimation with the above panel linear model is done by differencing the model at $t=3$ so that it becomes a cross-section model with $Q$ as a binary 'treatment':

$$\Delta Y_{i3} = \Delta\beta_3 + \beta_d Q_i + \beta_x' \Delta X_{i3} + \Delta U_{i3}$$

because $Q_i 1[t=3] = Q_i$ at $t=3$ and 0 at $t=2$, and $C_i$ in $W_{it}$ drops out to leave only $\Delta X_{i3}$ in $\Delta W_{i3}$. Under $Cor(\Delta U_3, Q) = Cor(\Delta U_3, \Delta X_3) = 0$, the OLS to the $\Delta Y_3$ model is consistent for $(\Delta\beta_3, \beta_d, \beta_x)$, but non-zero correlations can be

allowed as follows (Lee, 2019); recall our discussion right after (2.1).

Define the linear projection $L(Y \mid Z)$ of $Y$ on $Z$ and its residual $R(Y \mid Z)$ as

$$L(Y \mid Z) \equiv E(YZ')E^{-1}(ZZ')Z \quad \text{and} \quad R(Y \mid Z) \equiv Y - L(Y \mid Z)$$

where $E^{-1}(\cdot)$ stands for inverse. With $Z = (1, \Delta X_3')'$, rewrite the above $\Delta Y_3$ equation as

$$\Delta Y_3 = (\Delta\beta_3, \beta_x')Z + \beta_d Q + \Delta U_3$$

to obtain $L(\Delta Y_3 \mid 1, \Delta X_3)$ and $R(\Delta Y_3 \mid 1, \Delta X_3)$:

$$\begin{aligned} & L(\Delta Y_3 \mid 1, \Delta X_3) = (\Delta\beta_3, \beta_x')Z + \beta_d L(Q \mid 1, \Delta X_3) + L(\Delta U_3 \mid 1, \Delta X_3) \\ \Rightarrow\ & R(\Delta Y_3 \mid 1, \Delta X_3) = \beta_d R(Q \mid 1, \Delta X_3) + R(\Delta U_3 \mid 1, \Delta X_3) \\ & \text{(as } (\Delta\beta_3, \beta_x')Z \text{ drops out).} \end{aligned}$$

This "partial linear regression model" shows that $Cor\{R(Q \mid 1, \Delta X_3), R(\Delta U_3 \mid 1, \Delta X_3)\} = 0$, which allows $Q$ and $\Delta U$ to be related through $\Delta X_3$, is enough for the OLS to the $\Delta Y_3$ model to be consistent for $\beta_d$, although not necessarily for $(\Delta\beta_3, \beta_x)$.

As an example, we look at effects of "Tayo" bus (Hwang and Lee, 2018) which is just an usual bus with a makeover using cute animation characteristics on its exterior. In April and May of 2014, Seoul introduced 4 and 100 Tayo buses, respectively, along some bus stops. Seoul has 25 districts, each district has about 17 'dongs', and the overall number of bus stops in Seoul is about 8070.

The treatment in this study is not just 'whether a bus stop gets any Tayo bus route $(Q)$' times April/May dummy, but the number of Tayo bus routes that go through the bus stop, which is the *treatment dose/intensity*; see Campbell and Brakewood (2017) for a related example with the number of bike docks as the treatment (bike-sharing) intensity. A similar situation occurs with a minimum wage increase: when a minimum wage goes up, industries are affected differently, depending on the gap between the industry average wage and the new minimum wage, and the gap reveals the treatment intensity.

Since only dong-level aggregate variables are available, dividing the dong-level variables by the number of the bus stops in the dong, an 'averaged dong model per bus stop' is obtained. Table 'OLS for Differenced Monthly Rider Number' shows that one more Tayo bus route at a bus stop increases its riders there by 54 in April (insignificant) and 103 in May (significant) with some covariates including the number of women at each dong controlled. Since the dong-level correlation

between the number of women and the number of men is 0.99 in the data, controlling the number of women is essentially the same as controlling the dong population size. Because it costs only about \$1000 for a Tayo bus makeover and because the bus fare was about \$1 per trip within the city boundary, the table shows that the Tayo bus policy is a highly cost-effective treatment.

| OLS for Differenced Monthly Rider Number with $N = 423$ 'Dongs' | | |
|---|---|---|
| | April-March $\hat{\beta}_d$ (tvc, tv) | May-March $\hat{\beta}_d$ (tvc, tv) |
| # *Tayo routes* | *54.0 (0.31, 0.38)* | *103 (1.72, 2.24)* |
| # women | 15.3 (4.5, 5.3) | 17.4 (8.4, 8.5) |
| tvc, t-value for dongs clustered; tv, t-value; $R^2 = 0.33$ & 0.37 | | |

Instead of applying OLS to the differenced linear model which controls covariates parametrically, we can apply other estimators that control $\Delta X$ semi- or non-parametrically. Among those estimators, Lee's (2018) simple OLS-probit-based estimator for the effect on the population performs best, nearly dominating the other alternatives such as propensity matching, regression imputation and (inverse probability) weighting estimators.

For the effect on the treated, Abadie (2005) proposed a weighting estimator that is a sample analog of

$$E\left\{\frac{\Delta Y}{P(Q=1)}\frac{Q-P(Q=1\,|\,\Delta X)}{1-P(Q=1\,|\,\Delta X)}\right\},$$

replacing $P(Q=1\,|\,\Delta X)$ with probit/logit and $P(Q=1)$ with the T group proportion. Weighting estimators, however, tend to be numerically unstable due to the problem of denominators close to zero.

Both Lee (2018) and Abadie's (2005) estimators specify only the $Q$ equation, but not the $\Delta Y$ equation. This suggests that, although we conditioned only on $\Delta X$ because $\Delta Y$ depends only on $\Delta X$ in the linear model, we may want to condition on $W_2^3$ more extensively for Lee (2018) and Abadie's (2005) estimators, in case the linear model does not hold.

For more than two waves, consider a panel linear model:

$$Y_{it} = \beta_t + \beta_q Q_i + \beta_d Q_i 1[\tau \le t] + \beta_w' W_{it} + \delta_i + U_{it}, \quad t = 0,\ldots,T; \qquad (M_1)$$

the usual panel 'random-effect' estimators can be applied to this. Now, observe

$$\Delta\beta_t = \Delta\beta_1 + (\Delta\beta_2 - \Delta\beta_1)1[t=2] + , \ldots, + (\Delta\beta_T - \Delta\beta_1)1[t=T].$$

First-difference $M_1$ to obtain, as $\Delta Q_i 1[\tau \le t] = Q_i 1[t = \tau]$,

$$\Delta Y_{it} = \Delta\beta_1 + (\Delta\beta_2 - \Delta\beta_1)1[t = 2] + , \ldots , + (\Delta\beta_T - \Delta\beta_1)1[t = T] + \beta_d Q_i 1[t = \tau] + \beta_x' \Delta X_{it} + \Delta U_{it}. \quad (M_1')$$

The usual panel 'fixed-effect' estimators can be applied to this differenced model.

Alternatively to using panel data estimators, we can average the before and after periods to get only two averaged periods around $\tau$, which can be then differenced to obtain a single cross-section model. This may sound too "elementary" an approach, but as Bertrand et al. (2004) argued, this could be a robust way to conduct DD inference when many panel waves are highly correlated.

As an empirical example for DD with more-than-two-wave panel, consider effects of daylight saving time (DST) which is intended to save energy by starting working early when there is more light. Clocks are moved forward in the spring (and then back); see Choi et al. (2017) and references therein for DST in general. The Northeastern counties in Indiana adopted DST in 2006, with the other counties having DST already. The Northeastern counties constitute the T group, and the other counties the C group. Kotchen and Grant (2011) used panel data with $Y_{it} = \ln$(average daily residential electricity consumption in kilowatt hours) and Table 'Daylight Saving Time on Electricity' shows that DST increased electricity usage at home, contrary to the goal of DST. This empirical example is unusual, because the C group is always treated, as opposed to never treated. This is 'DD in reverse' (Kim and Lee, 2019), to be examined in detail later.

| Daylight Saving Time on Electricity (no covariate controlled) | | | |
|---|---|---|---|
| | 2004∼2005 (before) | 2006 (after) | BA |
| T group | 3.1256 | 3.1814 | 0.0558 |
| C group | 3.2239 | 3.2607 | 0.0368 |
| DD | ($N$ = 384,083) | | 0.0190 |

## 2.3. Generalizations of Panel Linear Model

If the treatment effect varies over time or if the treatment timing varies across individuals, then we may use

$$Y_{it} = \beta_t + \beta_q Q_i + \sum_{a=0}^{T-\tau_i} \beta_{da} Q_i 1[t = \tau_i + a] + \beta_w' W_{it} + \delta_i + U_{it} \quad (2.2)$$

where $\tau_i$ is the treatment timing for individual $i$, and $\beta_{d0}$ is the treatment effect of getting treated in the same period, $\beta_{d1}$ is the treatment effect of getting treated one period ago, etc. The covariates $W_{it}$ may interact with $Q_i$ so that

$Q_i W_{it}$ appears extra in the model, and $W_{it}$ may interact also with the treatment $Q_i 1[t = \tau_i + a]$ so that $W_{it} Q_i 1[t = \tau_i + a]$ appears as well, although we omit these generalizations.

We call the general case just examined '*varying effect (across time) and varying timing (across individuals)*'. In view of this, we can also think of models with constant effect ($\beta_d$) and constant timing ($\tau$)—the basic DD model—constant effect ($\beta_d$) and varying timing ($\tau_i$), and varying effect ($\beta_{d0}, \beta_{d1}, \ldots$) and constant timing ($\tau$). All these special cases can be handled by the above varying-effect and varying-timing model.

One caution is that there are two senses in which treatment effect is time-varying: calendar-time effect and duration effect. The former is that the effect differs depending on when the treatment takes place, and the latter is that the effect differs depending on how long ago the treatment was administered. Our time-varying effect is the duration effect, not the calendar-time effect; we may accommodate both (with $\beta_{da,t}$), but for simplicity, we entertain only the duration effect.

An interesting question is what happens if we use the constant effect model although the effect is actually time-varying. For a simple model

$$Y_{it} = \beta_1 + \sum_{a=0}^{T-\tau} \beta_{da} Q_i 1[t = \tau + a] + U_{it},$$

if we use its constant-effect version with $\beta_d Q_i 1[\tau \le t]$ replacing $\sum_{a=0}^{T-\tau} \beta_{da} Q_i 1[t = \tau + a]$, then it can be shown that the panel OLS for $\beta_d$ is consistent for the average of all effects, $(T - \tau + 1)^{-1}(\beta_{d0} + \beta_{d1} +, \ldots, +\beta_{d,T-\tau})$. For more general models, our conjecture is that we get to estimate a weighted average of the time-varying effects.

Related to this issue, instead of the "lasting" treatment $D_{it} = Q_i 1[\tau \le t]$, consider the 'one-shot/one-off' treatment $D_{it} = Q_i 1[\tau = t]$ which is applied only at one period to be withdrawn next; e.g., Japan gave out shopping coupons in 1999 (Hsieh et al., 2010), and the Taiwanese government gave out shopping vouchers in 2009 to stimulate the economy in the wake of the 2008 financial crisis (Kan et al., 2017). Unless otherwise mentioned, however, we will stick to lasting treatment. One reason for this is simplicity, and another reason is that even if the treatment is one-off, still its effect can be well found using (2.2) allowing time-varying effects. If we use $D_{it} = Q_i 1[\tau \le t]$ for one-shot treatment, however, the effect will come out almost zero, because we get to average all future effects where only one effect (or a few right after $\tau$) is non-zero.

Recalling the minimum wage example with treatment dose, suppose that industry $j$ is affected by an observed dose $\lambda_j$. One model for this is (covariates omitted), with '$\text{wage}_{min}$' for minimum wage,

$$Y_{it} = \beta_t + \sum_{j=1}^{J} \beta_{qj} Q_{ij} + \beta_d \sum_{j=1}^{J} \lambda_j Q_{ij} 1[\tau \le t] + \delta_i + U_{it},$$
$$\lambda_j \equiv 1[\text{wage}_{\min} > \text{wage}_j](\text{wage}_{\min} - \text{wage}_j)$$

where $Q_{ij} \equiv 1[\,i \text{ in industry } j\,]$, $\beta_{qj} Q_{ij}$ is the intercept-shift in industry $j$, and industry 0 with $\lambda_0 = 0$ serves as the base industry. Usually in DD, we consider only a common intercept shift by $\beta_q Q_i$ for the T group as in (2.2), but this model allows different intercept shifts $\beta_{qj}$'s across all industries. The treatment dose $\lambda_j$ is similar to Card and Krueger (1994, p.779) who used $\lambda_j / \text{wage}_j$ instead of $\lambda_j$.

The preceding model is a special case with $\beta_{pj} = 0$ for all $j$ of a more general model:

$$Y_{it} = \beta_t + \sum_j \beta_{qj} Q_{ij} + \sum_j \beta_{pj} 1[\text{wage}_{\min} > \text{wage}_j] Q_{ij} 1[\tau \le t]$$
$$+ \beta_d \sum_j \lambda_j Q_{ij} 1[\tau \le t] + \delta_i + U_{it}$$

where $\beta_{pj}$ is the intercept-shifting (or constant) part of the treatment effect for industry $j$, and $\beta_d \lambda_j$ is the effect proportional to the treatment dose $\lambda_j$. The mere fact of getting treated might have an effect $\beta_{pj}$, and $\beta_d \lambda_j$ is the extra effect depending on the level of $\lambda_j$.

South Korea raised its minimum wage by about 17% in January 2018. Its effect on industry $j$ $(Q_j = 1)$ at month $t$ may be estimated by a BA such as

$$E(Y_{18,t} \mid Q_j = 1) - E(Y_{17,Dec} \mid Q_j = 1) = E(Y^1_{18,t} \mid Q_j = 1) - E(Y^0_{17,Dec} \mid Q_j = 1)$$

for outcome $Y$ (employment, price level, etc.) and month $t$. Rewrite this BA as

$$E(Y^1_{18,t} \mid Q_j = 1) - E(Y^0_{18,t} \mid Q_j = 1) + E(Y^0_{18,t} \mid Q_j = 1) - E(Y^0_{17,Dec} \mid Q_j = 1)$$
$$= E(Y^1_{18,t} \mid Q_j = 1) - E(Y^0_{18,t} \mid Q_j = 1) \quad \text{if} \quad E(Y^0_{18,t} \mid Q_j = 1) = E(Y^0_{17,Dec} \mid Q_j = 1).$$

This BA ID condition is, however, not plausible due to the economy expanding/shrinking, or due to the monthly variation between month $t$ and December. The following form of DD addresses this concern.

Omitting '$Q_j = 1$' for simplicity, we can do double differencing in the time domain:

$$E(Y_{18,t}) - E(Y_{17,Dec}) - \{E(Y_{17,t}) - E(Y_{16,Dec})\}$$

$$
\begin{aligned}
&= E(Y^1_{18,t}) - E(Y^0_{17,Dec}) - \{E(Y^0_{17,t}) - E(Y^0_{16,Dec})\} \\
&= E(Y^1_{18,t}) - E(Y^0_{17,Dec}) - \{E(Y^0_{18,t}) - E(Y^0_{17,Dec})\} \\
&\quad + [E(Y^0_{18,t}) - E(Y^0_{17,Dec}) - \{E(Y^0_{17,t}) - E(Y^0_{16,Dec})\}] \\
&= E(Y^1_{18,t}) - E(Y^0_{18,t}) \quad \text{under} \quad E(Y^0_{18,t}) - E(Y^0_{17,Dec}) = E(Y^0_{17,t}) - E(Y^0_{16,Dec}) .
\end{aligned}
$$

Unlike the BA ID condition $E(Y^0_{18,t}) = E(Y^0_{17,Dec})$, this DD ID condition allows $E(Y^0_{18,t}) - E(Y^0_{17,Dec}) \neq 0$ as long as the difference stays the same as one year before, which is an example for $\mathrm{ID}_{DDt}$ that appeared earlier.

Differently from the usual DD with "two cross-section group-wise BA's" for $Q = 1, 0,$ here the DD consists of two time-wise BA's for the single group $Q_j = 1$; hence, call this "*time-wise DD*". Later when TD is examined, similar forms of TD will be seen: TD with two cross-section group-wise DD's, and TD with two time-wise DD's for a single group.

## III. DD with Repeated Cross-Sections (RCS)

This section examines DD with RCS. Our ID and estimation discussion for RCS is relatively brief, because the main points were seen already with panel data. Some topics relevant to both panel data and RCS such as limited dependent variable models and "fuzzy DD" below, but left out in the preceding panel section due to the length concern, are also examined in this section.

In RCS, an individual is observed only once. A person may be observed more than once, but the possibility is slim and thus ignored. With $t = 2, 3$, Define the $t = 3$ sampling dummy:

$$S_i = 1[\text{individual } i \text{ sampled at } t = 3];$$

where '$S$' is from 'sampled'. What is observed for RCS with $t = 2, 3$ is

$$Q_i, \quad S_i, \quad W_i = (1 - S_i) W_{i2} + S_i W_{i3} \quad \text{and} \quad Y_i = (1 - S_i) Y_{i2} + S_i Y_{i3} .$$

For more than two periods, let $S_i$ denote the sampled period and $S_{it} \equiv 1[S_i = t]$ to have

$$W_i \equiv \sum_t W_{it} S_{it}, \quad Y_i \equiv \sum_t Y_{it} S_{it} \quad \text{and} \quad U_i \equiv \sum_t U_{it} S_{it} .$$

Assume that $S$ is independent of all potential responses, $Q$ and $W$. This aspect

—the relationship between $S$ and the other random variables—does not arise in panel data, although unbalancedness in panel data is an analogous problem.

### 3.1. Identification with RCS

DD with RCS is

$$\begin{aligned} DD_{23} &\equiv E(Y \mid Q=1,S=1)-E(Y \mid Q=1,S=0) \\ &\quad -\{E(Y \mid Q=0,S=1)-E(Y \mid Q=0,S=0)\} \\ &= E(Y_3^1 \mid Q=1)-E(Y_2^0 \mid Q=1)-\{E(Y_3^0 \mid Q=0)-E(Y_2^0 \mid Q=0)\}. \end{aligned}$$

Then

$$\begin{aligned} &DD_{23} \equiv E(Y_3^1 - Y_3^0 \mid Q=1) \quad \text{under} \\ &E(Y_3^0 \mid Q=1)-E(Y_2^0 \mid Q=1)=E(Y_3^0 \mid Q=0)-E(Y_2^0 \mid Q=0); \end{aligned} \qquad (\mathrm{ID}_{4D})$$

we write "$\mathrm{ID}_{4D}$", as 4 groups are involved. The right-hand side of $\mathrm{ID}_{4D}$, which is the second part of $DD_{23}$, is identified: if it is zero, use just the first half of the DD.

Put together the same-period expected values in $\mathrm{ID}_{4D}$ on the same side:

$$E(Y_3^0 \mid Q=1)-E(Y_3^0 \mid Q=0)=E(Y_2^0 \mid Q=1)-E(Y_2^0 \mid Q=0). \qquad (\mathrm{ID}'_{4D})$$

$\mathrm{ID}'_{4D}$ is a 'stationarity' condition, because the effect of $Q$ on $Y_3^0$ at $t=3$ is the same as the effect of $Q$ on $Y_2^0$ at $t=2$. The right-hand side of $\mathrm{ID}'_{4D}$ is also identified, and if it is zero, use just the cross-sectional group difference at $t=3$ instead of DD. Which part of DD might be redundant can be checked out to reduce a DD to a BA. A caution is "$E(Y_3^0 \mid Q=1)-E(Y_2^0 \mid Q=1) \neq E(Y_3^0 - Y_2^0 \mid Q=1)$", because what actually appears with $W_t$ in is $E(Y_3^0 \mid W_3=w, Q=1)-E(Y_2^0 \mid W_2=w, Q=1)$, which cannot be merged.

For the effect on the untreated, it can be easily shown that

$$\begin{aligned} &DD_{23} = E(Y_3^1 - Y_3^0 \mid Q=0) \quad \text{under} \\ &E(Y_3^0 \mid Q=1)-E(Y_2^0 \mid Q=1)=E(Y_3^1 \mid Q=0)-E(Y_2^0 \mid Q=0). \end{aligned} \qquad (\mathrm{ID}''_{4D})$$

Under $\mathrm{ID}_{4D}$ and $\mathrm{ID}''_{4D}$, $DD_{23} = E(Y_3^1 - Y_3^0)$.

Suppose $Y$ is a limited dependent variable (LDV) based on a latent continuous $Y^*$ whose model is linear. For this, the DD ID findings hold for $Y^*$, and $\beta$ for $Y^*$ can be estimated with probit, tobit, etc., depending on the nature of the LDV. That is, the treatment effect on $Y^*$ is the slope of $QS$, which gives the DD

interpretation in terms of $E(Y^*|\cdot)$, but the DD interpretation in terms of $E(Y|\cdot)$ no longer holds; see, e.g., Puhani (2012) and Kim and Lee (2017).

If $Y \geq 0$ as in count response which is not based on any latent response $Y^*$, we can use the popular exponential specification:

$$E(Y|Q,S,W) = \exp(\beta_1 + \beta_\tau S + \beta_q Q + \beta_d QS + W'\beta_w) .$$

Here $\beta_d$ can be interpreted as a "ratio in ratios (RR)" effect—ratio is analogous to BA:

$$\left(\frac{\exp(\beta_1 + \beta_\tau + \beta_q + \beta_d + w'\beta_w)}{\exp(\beta_1 + \beta_q + w'\beta_w)}\right) / \left(\frac{\exp(\beta_1 + \beta_\tau + w'\beta_w)}{\exp(\beta_1 + w'\beta_w)}\right) = e^{\beta_d} :$$

$D = QS$ increases $E(Y|Q,S,W)$ by $\exp(\beta_d)$ times.

In the linear model, the group effect $\beta_q$ and time effect $\beta_\tau$ are removed by DD to result in $DD = \beta_d$. Analogously, with an exponential model, the group and time effects are removed by RR to result in $RR = \exp(\beta_d)$. This is no surprise when we use the exponential model, but if we proceed only with conditional means, then the following may be surprising:

$$\left(\frac{E(Y|W,Q=1,S=1)}{E(Y|W,Q=1,S=0)}\right) / \left(\frac{E(Y|W,Q=0,S=1)}{E(Y|W,Q=0,S=0)}\right) = e^{\beta_d} .$$

In South Korea, platform screen doors (PSD) in subways were installed to prevent suicides. Using the monthly suicide number panel data over 2003-2012 at 121 subway stations of Seoul Metro, which is one of the Seoul subway companies, Chung et al. (2016) estimated the effect. Seoul Metro installed PSD's over 2005-2009, which gave a DD framework with different treatment timings across different stations. There were two types of PSD: 'full PSD' extending from floor to ceiling, and 'half PSD' extending chest-high at 1.65m. Poisson regression gave the slope of PSD dummy −2.2 with 95% confidence interval (CI) (−3.5, −0.84): PSD reduced subway suicides by $100\{1-\exp(-2.2)\} = 89\%$. When separate dummies for full PSD and half PSD were used, the slopes were, respectively, −17 with CI (−18, −16) and 0.75 with CI (−1.13, 2.62): full PSD eradicated subway suicides, but half PSD was useless.

### 3.2. Estimation with RCS

With treatment applied at $\tau$ onwards, consider a constant-effect and constant-

timing panel linear model without the individual-specific effect $\delta_i$ :

$$Y_{it} = \beta_t + \beta_q Q_i + \beta_d Q_i 1[\tau \le t] + W_{it}'\beta_w + U_{it}.$$

The corresponding RCS model is, with $D_i = Q_i 1[\tau \le S_i]$ and $S_{it} \equiv 1[S_i = t]$,

$$Y_i = \beta_1 + \Delta\beta_{21} S_{i2} + , \ldots, + \Delta\beta_{T1} S_{iT} + \beta_q Q_i + \beta_d D_i + W_i'\beta_w + U_i$$

where the sampling dummies capture the time effects, and individual $i$ is treated $(D_i = Q_i 1[\tau \le S_i] = 1)$ if he/she is sampled at or after $\tau$ with $Q_i = 1$. We can do the OLS of $Y$ on $(1, S_2, \ldots, S_T, Q, D, W)$.

For the basic two period constant-effect and constant-timing case, instead of the OLS, we may use a weighting estimator for the effect on the treated in Abadie (2005) that is a sample analog for

$$E\left[\frac{P(Q=1|W)}{P(Q=1)} \cdot \frac{S - E(S)}{E(S)\{1-E(S)\}} \frac{Q - P(Q=1|W)}{P(Q=1|W)P(Q=0|W)} \cdot Y\right].$$

As was already noted, weighting estimators tend to be numerically unstable though.

A generalization of the constant-effect and constant-timing panel linear model is a varying-effect and varying-timing panel linear model:

$$Y_{it} = \beta_t + \beta_q Q_i + \sum_{a=0}^{T-\tau_i} \beta_{da} Q_i 1[t = \tau_i + a] + W_{it}'\beta_w + U_{it}.$$

For the RCS model derived from this, we can do the OLS of $Y_i$ on

$$1, S_{i2}, \ldots, S_{iT}, \ Q_i, \ \tilde{D}_{i0}, \tilde{D}_{i1}, \ldots, \tilde{D}_{i,T-1}, \ W_i \ \text{ with } \ \tilde{D}_{ia} \equiv Q_i 1[S_i = \tau_i + a]$$

where the treatment effect varies, depending on when the individual is sampled; $1[S_i = \tau_i + a] = 1$ means 'sampled $a$ periods after $\tau_i$'. Clearly, this includes 'constant-effect and varying-timing' and 'varying-effect and constant-timing' as special cases.

As an example for varying time-effect and constant-timing, Eissa and Liebman (1996) examined the effect of earned income tax credit (EITC) on work or not (binary $Y$); EITC reduces income tax, and is applied only to single women with low income and at least one children. The Current Population Survey data for 1984-1986 (before) and 1988-1990 (after) were used, which is a RCS, where the T group is EITC-eligible with children and the C group is EITC-ineligible due to no

children although their income is low enough.

Table 'Probit with Marginal Effect in $\{\cdot\}$ for EITC' is part of the results in Eissa and Liebman (1996). When effect constancy is assumed in the left column, the estimated marginal effect (i.e., the effect on $P(Y=1|W)$) is 0.019, which is more or less the average of the three marginal effects (0.008, 0.029 and 0.028) on the right column where the effect is allowed to vary over time. The initial period effect is small (0.008), perhaps due to the lack of policy awareness in the beginning. Unawareness of a policy despite its existence does happen; see, e.g., Kim et al. (2012) and Kuo (2012).

| Probit with Marginal Effect in $\{\cdot\}$ for EITC (SE in $(\cdot)$) | | |
|---|---|---|
| | Time-Constant Effect | Time-Varying Effect |
| Kids ($Q$) | -0.250 (0.029) | -1.462 (0.110) |
| Post86 | 0.019 (0.031) | |
| Kids × Post86 | *0.074 (0.030) {0.019 (0.008)}* | |
| Kids × 1988 | | *0.033 (0.057) {0.008 (0.014)}* |
| Kids × 1989 | | *0.116 (0.058) {0.029 (0.015)}* |
| Kids × 1990 | | *0.112 (0.057) {0.028 (0.015)}* |

Hagiwara et al. (2013) examined effects of a maternal child health care book on health-related behaviors in Palestine. The book has 56 pages, and contains helpful health information and the health records of mother and children before and after birth. The data are a RCS collected Jan.-Feb., 2007 (before period) and March-April 2008 (after): 260 and 270 women in the T group before and after, and 70 and 70 women in the C group. An interesting point is that the T group is from 24 centers and the C group is from 6 centers, where the centers were randomly selected from 49 health centers. This gave a panel feature to the data (in addition to a randomization feature), because some women are from the same center, which is then taken care of by including the center dummies in estimation.

Part of Table 2 in Hagiwara et al. (2013) is Table 'Effect of Health Book on Health Behaviors'; the response variables other than Center Hours are binary to which OLS was applied as well. Column 'Appoint' for being aware of the next appointment shows that, although the treatment itself is not significant, its interaction with primary education dummy significantly increases the awareness. Column 'Center Hours' for the hours spent in the center shows that, although 1st delivery itself has a significant negative effect, its interaction with $D$ has a significantly positive effect (26 hours). In Column 'On Breast Feed' for knowledge on breast feeding, the interaction between $D$ and being literate increases the knowledge probability by 32%. In Column 'On Rupture' for knowledge on membrane rupture, $D$ increases the knowledge probability by 20%.

| Effect of Health Book on Health Behaviors: Estimate (t-value) | | | | |
|---|---|---|---|---|
| | Appoint | Center Hours | On Breast Feed | On Rupture |
| $D$ | 0.028 (0.66) | -6.1 (0.64) | -0.12 (1.1) | 0.20 (2.0) |
| $D \times$ literate | 0.11 (1.4) | -4.8 (0.26) | 0.32 (1.9) | 0.083 (0.63) |
| $D \times$ primary | 0.083 (2.3) | 8.6 (0.83) | -0.009 (0.08) | -0.10 (1.1) |
| $D \times$ 1st delivery | -0.16 (1.5) | 26 (2.1) | 0.017 (0.11) | -0.11 (0.83) |
| 1st delivery | 0.077 (1.8) | -35 (3.2) | -0.056 (0.44) | -0.02 (0.21) |
| Appoint, aware of next appointment; Center Hours, hours in center; primary, primary edu. On Breast Feed, breast feeding knowledge; On Rupture, membrane rupture knowledge | | | | |

### 3.3. Fuzzy DD

Sometimes $D_{it} \neq Q_i 1[\tau \leq t]$ happens, which may be called "fuzzy DD", relative to the usual "sharp DD" with $D_{it} = Q_i 1[\tau \leq t]$. We can then use $Q_i 1[\tau \leq t]$ as an instrument for $D_{it}$, if $Q_i 1[\tau \leq t]$ is plausibly excluded from the $Y_{it}$ model. The terminology "fuzzy DD" and "sharp DD" are taken from the 'fuzzy RD' and 'sharp RD' in the RD literature, and they were used first in Lee (2016a). De Chaisemartin and D'Haultfoeuille (2018) used the same terminology later in showing that the IVE estimates a local (weighted) average treatment effect under certain assumptions.

In 1959, Norway decided to increase the mandatory schooling years from 7 to 9: students should remain at school until age 16 as they start schooling at 7. All municipalities were mandated to implement the reform by 1973. As the consequence, the municipalities have different reform years over 1960∼1973; $D$ is schooling years (not binary) in this example. Whether a woman was affected by the reform or not was determined by her age (say, $1[\tau \leq t]$) and the municipality of residence (say, $Q$). The chances for individuals moving across municipalities during the periods are thought to be very low, and thus ignored.

The first cohort possibly subject to the reform was those born in 1947, because they were supposed to finish primary school in 1961=1947+14. The last cohort that might have not experienced the reform was those born in 1958, because they could have completed 7 year compulsory schooling by 1972=1958+14. Monstadt et al. (2008) observed all women born between 1947 and 1958 until 2002. In 2002, the youngest were 44 as they were born in 1958, which means that almost all women in the data had completed fertility. The model in Monstadt et al. (2008) is

$$Y_i = \beta_1 + \sum_{j=1948}^{1958} \beta_{cj} 1[i \text{ born in year } j] + \sum_{j=2}^{672} \beta_{mj} 1[i \text{ in municipality } j] + \beta_d D_i + U_i$$

and IVE is applied with schooling $D_i$ instrumented by $1[i$ subject to reform$]$.

Part of Table 2 in Monstadt et al. is ('*' for significance at 5% level)

| $Y$ | #children | 1st birth in 15-20 | 1st birth in 20-25 | 1st birth in 35-40 |
|---|---|---|---|---|
| OLS | -0.013 (0.004)* | -0.032 (0.001)* | -0.024 (0.001)* | 0.005 (0.000)* |
| IVE | -0.009 (0.087) | -0.080 (0.039)* | 0.044 (0.032) | 0.021 (0.009)* |

While OLS indicates a significant decrease by 0.013 in #children, IVE does not. IVE shows 8% decrease in the first birth in ages 15-20, and 2.1% increase in ages 35-40: women postpone births due to more schooling, but schooling does not affect the overall fertility. This is one of a few recent studies showing no effect of education on fertility to challenge the conventional "wisdom" that education decreases fertility; see Kan and Lee (2018) and references therein.

## IV. DD in Reverse (DDR)

In some DD, the control group is always treated, instead of already untreated, which is called "DD in reverse (DDR)". Here, we examine DDR based on Kim and Lee (2019). One example for DDR is building a bridge for a region across a river, along which other regions already have bridges (Mahmud and Sawada, 2018), and more examples can be seen in Chemin and Wasmer (2009) and Kotchen and Grant (2011). More generally, when the treatment timing $\tau_i$ varies across individuals, DDR occurs at $\max_{i=1,...,N} \tau_i$ because all the other individuals have been treated. Since the C group is always treated in DDR, it is awkward to call the $Q=1$ group the 'T group'; instead, call them the 'switching group'.

DDR takes the same form as DD, but the potential responses differ:

$$\begin{aligned} DDR_{23} &\equiv E(\Delta Y \mid Q=1) - E(\Delta Y \mid Q=0) \\ &= E(Y_3^1 \mid Q=1) - E(Y_2^0 \mid Q=1) - E(\Delta Y_3^1 \mid Q=0)\,. \end{aligned}$$

Subtract and add the counterfactual $E(Y_3^1 \mid Q=1) - E(Y_2^1 \mid Q=1) = E(\Delta Y_3^1 \mid Q=1)$ after the second term to get

$$\begin{aligned} DDR_{23} &= E(Y_3^1 \mid Q=1) - E(Y_2^0 \mid Q=1) - \{E(Y_3^1 \mid Q=1) - E(Y_2^1 \mid Q=1)\} \\ &\quad + E(\Delta Y_3^1 \mid Q=1) - E(\Delta Y_3^1 \mid Q=0) = E(Y_2^1 - Y_2^0 \mid Q=1) \quad \text{under} \\ &E(\Delta Y_3^1 \mid Q=1) = E(\Delta Y_3^1 \mid Q=0)\,. \end{aligned} \qquad (\mathrm{ID}_{DDR})$$

*DDR is the effect on the "switched" ($Q=1$) at the pre-switch period* $t=2$, differently from DD identifying the effect on the $Q=1$ group at the post-treatment period $t=3$.

Consider a panel data model for $t = 2,3$ with $D_{it} = (1-Q_i) + Q_i 1[t=3]$:

$$Y_{it} = \beta_t + \beta_q Q_i + \beta_d (1 - Q_i + Q_i 1[t=3]) + W_{it}'\beta_w + U_{it}.$$

This gives a RCS model for OLS: recalling $S_i \equiv 1[\,i$ sampled at $t=3\,]$,

$$\begin{aligned} Y_i &= \beta_2 + (\beta_3 - \beta_2) S_i + \beta_q Q_i + \beta_d (1 - Q_i + Q_i S_i) + W_i'\beta_w + U_i \\ &= (\beta_2 + \beta_d) + (\beta_3 - \beta_2) S_i + (\beta_q - \beta_d) Q_i + \beta_d Q_i S_i + W_i'\beta_w + U_i : \end{aligned}$$

$\beta_d$ is still estimated by $QS$ as in DD. The difference from DD is that the slope of $Q$ in DDR is $\beta_q - \beta_d$, whereas it is $\beta_q$ in DD. If $(Q, 1-Q-QS)$ is used as regressors instead of $(Q, QS)$ in DDR, then the slope of $Q$ becomes $\beta_q$.

Since both groups are treated, one may think of identifying the treatment effects separately, say $\beta_{d0}$ and $\beta_{d1}$, for the two groups $Q = 0,1$. This, however, does not work as only $\beta_{d1}$ is identified, which can be seen in the following generalized panel model:

$$\begin{aligned} Y_{it} &= \beta_t + \beta_q Q_i + \beta_{d0}(1 - Q_i) + \beta_{d1} Q_i 1[t=3] + W_{it}'\beta_w + U_{it} \\ &= (\beta_t + \beta_{d0}) + (\beta_q - \beta_{d0}) Q_i + \beta_{d1} Q_i 1[t=3] + W_{it}'\beta_w + U_{it}. \end{aligned}$$

As elementary as this may look, this demonstrates that the treatment effect is estimable only for those whose treatment status changes.

As an empirical example, the South Korean work hours have been reduced from 44 to 40 in different years, depending on firm size:

| $1000^+$ employees (big firms) | 300-999 (small firms) | 100-299 | 50-99 | 20-49 | 5-19 |
|---|---|---|---|---|---|
| 2004 ($t=2$) | 2005 ($t=3$) | 2006 | 2007 | 2008 | 2011 |

The enforcement of the law was lax, however, and thus it is unclear to what extent the law affected the actual work hours, as well as the real wage. DD for 2003-2004 provides the effect on big firms (treated) at 2004 (post-treatment period), whereas DDR for 2004-2005 provides the effect on small firms (switched) at 2004 (pre-switch period).

Using the Occupational Employment Statistics in South Korea, which is a RCS, Kim and Lee (2019) applied both DD and DDR and some of their findings is in Table 'DD and DDR Effects with No Covariate Controlled' which shows a reduction of about two hours and an increase in weekly real wage by about $50-70; both changes are statistically significant.

| DD and DDR Effects with No Covariate Controlled: Average (SE) | | | | |
|---|---|---|---|---|
| $Y$ : | Weekly work hours | | Weekly wage (in \$10) | |
| | Small Firms | Big Firms | Small Firms | Big Firms |
| 2003: | 52.88 | 52.66 | 63.11 | 71.55 |
| 2004: | 50.84 | 48.46 | 60.86 | 76.25 |
| 2004-2003: | -2.04 | -4.20 | -2.25 | 4.70 |
| DD for big firm firms 2004: | -4.20+2.04=-2.16 (0.45) | | 4.70+2.25=6.95 (1.37) | |
| DDR for small firms 2004: | -1.87 (0.44) | | 4.88 (1.41) | |

# V. Synthetic Control

Suppose there is a single treated individual and multiple possible controls, but none of the controls looks good on its own. Then, it is conceivable to linearly combine the controls to come up with an "artificial" control individual who is a good control in the sense that its "artificial past" is parallel to the untreated past of the treated individual. The artificial control is called a 'synthetic control'; see Abadie and Gardeazabal (2003), Abadie et al. (2010, 2015) and references therein.

## 5.1. Main Idea

Consider $J$ subjects $j=1,\ldots,J$ observed for periods $1,\ldots,T$, and only subject $J$ is treated at the last period $T$; otherwise, no treatment. The goal is to estimate $Y_{JT}^1 - Y_{JT}^0$, the effect for subject $J$ at the post-treatment period $T$, using a synthetic control $\sum_{j=1}^{J-1} w_j Y_{jT}$ as the counter-factual $Y_{JT}^0$, where the weight $w_j$'s satisfy $\sum_{j=1}^{J-1} w_j = 1$ and $w_j \geq 0$. Then the effect estimator is

$$Y_{JT} - \sum_{j=1}^{J-1} w_j Y_{jT} .$$

If there are multiple individuals in the T group, then we may use the T group aggregate as a single individual, or a synthetic control may be constructed separately for each individual in the T group; see, e.g., Kreif et al. (2016).

To find $w \equiv (w_1,\ldots,w_{J-1})'$, define pre-treatment variables

$$Z_j \equiv (Y_{j1},\ldots,Y_{j,T-1}, X_{j1}',\ldots,X_{j,T-1}')' .$$

For a given weighting matrix $V$, minimize with respect to (wrt) $w$

$$\left( Z_J - \sum_{j=1}^{J-1} w_j Z_j \right)' V \left( Z_J - \sum_{j=1}^{J-1} w_j Z_j \right).$$

If not for the constraints $\sum_{j=1}^{J-1} w_j = 1$ and $w_j \geq 0$, this is a OLS/GLS problem of predicting $Z_J$ with regressors $Z_1, \ldots, Z_{J-1}$.

As for choosing $V$, Abadie et al. (2010) chose $V$ as a p.d. diagonal matrix minimizing the prediction error only for $(Y_{J1}, \ldots, Y_{J,T-1})$ using the control group $Y$'s and weight $w(V)$ as follows: for a given $V$, $w(V)$ is obtained in the above OLS/GLS set-up, then $V$ is chosen by minimizing

$$\left\{ Y_{J,pre} - \sum_{j=1}^{J-1} w_j(V) Y_{j,pre} \right\}' \left\{ Y_{J,pre} - \sum_{j=1}^{J-1} w_j(V) Y_{j,pre} \right\} \quad \text{where} \quad Y_{J,pre} \equiv (Y_{j1}, \ldots, Y_{j,T-1})'.$$

This $V$-choice scheme is admittedly ad-hoc, but since the eventual goal is finding a synthetic control that satisfies the same time-effect assumption (i.e., parallel untreated paths in the past for the two groups), this way of choosing $V$ seems sensible, although restricting $V$ to a diagonal matrix looks still arbitrary.

As an empirical example, Abadie et al. (2010) analyzed the effect of California (CA) Proposition 99 in 1988 which increased the cigarette exercise tax by 25 cents per pack. State panel data over 1970-2000 where $Y_{it}$ is per capita cigarette consumption were used, with a donor pool of 38 states, excluding the states with other smoking-discouraging measures. The "smoking predictor" $Z_j$ in Abadie et al. (2010) includes the 1980-1988 averages of cigarette prices, logged per capita GDP, % population aged 15-24, per capita beer consumption, and $Y_{i,1975}$, $Y_{i,1980}$ and $Y_{i,1988}$. When the $w$ estimate was obtained, positive weights were only for Connecticut (0.164), Colorado (0.069), Montana (0.199), Nevada (0.234) and Utah (0.334). Abadie et al. (2010) concluded that smoking declined by 26 packs per capita per year due to the CA proposition 99, whereas a previous study found a decline of only 14 packs.

## 5.2. Inference with Permutation Test

Inference for synthetic control is done with 'permutation/randomization test'. To understand permutation test, consider 5 individuals $(Y_a, Y_b, Y_c, Y_d, Y_e)$ with $(Y_a, Y_b)$ in the T group, and the remaining three in the C group. Regard these $Y$ values as fixed, and think of assigning artificially the 5 individuals to the T and C groups. Obtain the actual treatment effect from the data with $(Y_a, Y_b)$ and $(Y_c, Y_d, Y_e)$ in the T and C groups, and a 'pseudo effect' from 'pseudo data' $(Y_d, Y_e)$ and $(Y_a, Y_b, Y_c)$ in the two groups:

$$m_{data} \equiv \frac{Y_a + Y_b}{2} - \frac{Y_c + Y_d + Y_e}{3} \quad \text{(actual)}, \quad m_j \equiv \frac{Y_d + Y_e}{2} - \frac{Y_a + Y_b + Y_c}{3} \quad \text{(pseudo)}.$$

Going further, obtain all possible pseudo effects, $m_1,\ldots,m_M$, by creating all possible pseudo data, and check how extreme $m_{data}$ is in the distribution of $m_1,\ldots,m_M$. For example, if $m_{data}$ is the 0.013 quantile in the distribution, then the p-value of the test is $0.026 = 2\times 0.013$ for the two-sided zero effect test: despite the zero true effect, having the estimated effect greater than $|m_{data}|$ in absolute magnitude is only 00026, which is the type I error probability of rejecting the correct null (zero effect) falsely.

Related to this cross-sectional "placebo test" is doing the same test with a pseudo $\tau$; this idea in fact applies to DD in general, not just to synthetic control, because such tests are often applied as part of sensitivity/robustness analysis. In the above CA proposition 99 example, we may set $\tau = 1980$, not 1988, and use only the left side of the original $\tau$ (e.g., 1970-1987) and repeat the same estimation and test procedure to see if a significant effect is found. If yes, something must be wrong because no actual treatment took place during 1970-1987.

Randomly setting $\tau$ raises the possibility to construct a pseudo effect distribution differently: instead of randomly assigning subjects, maintain the same subjects in the two groups, but randomly assign the treatment timing for the T group. This is fine under the null of no effect, which makes $\tau$ changeable. We may make the permutation test more elaborate by going both ways: randomly assign subjects to either group, and then randomly select $\tau$ in the T group. This increases the number of pseudo effect estimates, which is advantageous if the number of subjects is too small for one-dimensional permutation to generate enough pseudo effect estimates. Permutation test is also called 'randomization test' or 'Fisher's exact test'.

One problem in the permutation test is that permutation test requires 'exchangeability': e.g., for $(Y_a, Y_b)$, exchangeability is

$$P(Y_a \le y_1, Y_b \le y_2) = P(Y_b \le y_1, Y_a \le y_2) \quad \text{for all} \quad (y_1, y_2).$$

For $(Y_1,\ldots,Y_J)$, the probability should stay the same for any permutation of $(Y_1,\ldots,Y_J)$: e.g., $P(Y_1 \le y_1, Y_2 \le y_2,\ldots,Y_J \le y_J) = P(Y_J \le y_1, Y_{J-1} \le y_2,\ldots,Y_1 \le y_J)$ for all $(y_1,\ldots,y_J)$. Unfortunately, this is not plausible in most synthetic control applications; e.g., California (CA) would not be exchangeable with Montana. Also, if we randomize $\tau$ within California, exchangeability requires $(CA_1,\ldots,CA_T)$ to be exchangeable, which is not plausible either. Despite this problem, in synthetic control approach, there seems to be no other practical way to conduct inference than permutation test.

### 5.3. Applications and Remarks

As an empirical example of synthetic control, Abadie and Gardeazabal (2003) found that terrorism in Basque county reduced the per capita GDP of Basque county by 10 percentage point; here, other Spanish counties were used to construct the synthetic control. Abadie et al. (2015) examined effects of the 1990 German reunification using Austria, Japan, Netherlands, Switzerland and the U.S. for the synthetic control: the per capita GDP of Germany declined by $1600 per year, which is about 8% of the 1990 level.

Bohn et al. (2014) assessed the Arizona Legal Workers Act effect on the proportion of the Hispanic. They used, instead of $\bar{Y}_{J,post} - \bar{Y}_{synthetic,post}$,

$$\bar{Y}_{J,post} - \bar{Y}_{synthetic,post} - (\bar{Y}_{J,pre} - \bar{Y}_{synthetic,pre})$$

which allows an intercept difference between the two groups. They showed that the law decreased the non-citizen Hispanic population in Arizona.

Dupont et al. (2015) assessed the long-term effects of the Kobe earthquake in 1995 to find that the long term effects are localized and highly heterogeneous, because some adjacent areas actually benefitted from the disaster while the areas close to the city center suffered much. This study demonstrated the importance of looking at detailed data at a micro level, rather than examining only aggregate data at a higher level, which would mask the micro-level effect heterogeneity to suggest almost no effect as some studies have done.

Ando (2015) found that nuclear power facilities in Japan increased the local income by about 11%, Munasib and Rickman (2015) showed that shale oil and gas production in the U.S. had a positive effect on the local economy, but the effect was highly heterogeneous across different states. Xu (2017) found that allowing to register on election day in the U.S. instead of requiring to register before the election day increased voter turnout in states adopting the policy early.

Kreif et al. (2016) assessed the effect of 'hospital pay-for-performance' on mortality in the U.K. hospitals. They found an insignificant decrease in the mortalities for the incentivized categories including pneumonia and acute myocardial infarction, but a significant increase in the non-incentivised category mortality. This was in sharp contrast to the conventional DD result where a significant decrease and an insignificant increase were seen for the incentivized and non-incentivised categories, respectively. In Kreif et al. (2016), there were many treated units, which they aggregated to a single unit to apply the synthetic control method, while keeping the control units without aggregation.

As long as the goal is constructing the counterfactual untreated response for the treated at the post-treatment era, several issues arise in synthetic control approach

due to many restrictions. First of all, we may want to allow an intercept $w_0$ in the prediction error $Z_J - w_0 - \sum_{j=1}^{J-1} w_j Z_j$, because this can only improve predicting $Z_J$.

Second, we may want to allow $w_j$ to be negative because imposing $w_j \geq 0$ can only worsen the prediction error, although the motivation to insist on $w_j \geq 0$ (i.e., not using a control trending in the opposite direction) is sensible.

Third, we may do without the restriction $\sum_{j=1}^{J-1} w_j = 1$ because, again, the restriction can only worsen the prediction error, although the motivation to impose $\sum_{j=1}^{J-1} w_j = 1$ is understandable: to have a "full" individual, not more than full with $\sum_{j=1}^{J-1} w_j > 1$, nor less than full with $\sum_{j=1}^{J-1} w_j < 1$.

Fourth, instead of minimizing $|\text{T group past} - w \times (\text{C group past})|$ as in $Z_J - \sum_{j=1}^{J-1} w_j Z_j$, we may want to minimize $|\text{C group future} - \omega \times (\text{C group past})|$ wrt $\omega$. The pattern can be then taken to the T group past to construct the untreated T group future. That is, choose $\hat{\omega}$ minimizing the prediction error, and then construct the T group untreated future as $\hat{\omega} \times$ (T group past).

Doudchenko and Imbens (2016) compared various approaches imposing/relaxing the above constraints, and proposed penalizing a large number of positive weights and large magnitudes with $\sum_j |w_j|$, $\sum_j w_j^2$ or $\sum_j 1[w_j \neq 0]$, using "regularization/tuning" constants.

# VI. Triple Difference (TD) and More

TD appears if only a group of individuals (say, with $G=1$) among the $Q=1$ group are treated at $t=3$; call this '$TD_{23}$' with $t=2,3$ available. For example, $Q=1$ for an ethnic minority and $G=1$ for women, and an education program is applied only to women in the ethnic minority $(G=1, Q=1)$. This makes the TD treatment a *triple interaction* $D = GQ1[t=3]$, in contrast to the DD treatment $Q[t=3]$ that is only a double interaction. TD can be generalized to quadruple difference (QD) and beyond. QD needs another subgroup dummy, say $A$, so that only those with $AGQ1[t=3]=1$ get treated (quadruple interaction)—call this '$QD_{23}$'.

## 6.1. Identification for TD

Omitting $W_2^3 \equiv (C', X_2', X_3')'$, TD with panel data is

$$TD_{23} \equiv E(\Delta Y_3 \mid G=1, Q=1) - E(\Delta Y_3 \mid G=0, Q=1) \\ -\{E(\Delta Y_3 \mid G=1, Q=0) - E(\Delta Y_3 \mid G=0, Q=0)\}.$$

Then TD identifies the effect on the treated $(G=1, Q=1)$ at $t=3$:

$$TD_{23} = E(Y_3^1 - Y_3^0 \mid G=1, Q=1) \quad \text{under}$$
$$E(\Delta Y_3^0 \mid G=1, Q=1) - E(\Delta Y_3^0 \mid G=0, Q=1)$$
$$= E(\Delta Y_3^0 \mid G=1, Q=0) - E(\Delta Y_3^0 \mid G=0, Q=0). \qquad (\mathrm{ID}_{TD})$$

$\mathrm{ID}_{TD}$ is that the effect of $G$ on $\Delta Y_3^0$ is the same for $Q=1$ and $Q=0$. For an ethnic minority $(Q=1)$ education program applied only to females $(G=1)$, the left-hand side of $\mathrm{ID}_{TD}$ is the difference between the test score change for the minority females and that for the minority males, which is assumed to be the same as the difference between the test score change for the majority females and that for the majority males on the right-hand side.

$\mathrm{ID}_{TD}$ with $Q=1$ and $Q=0$ on the opposite sides can be rewritten as:

$$E(\Delta Y_3^0 \mid G=1, Q=1) - E(\Delta Y_3^0 \mid G=1, Q=0)$$
$$= E(\Delta Y_3^0 \mid G=0, Q=1) - E(\Delta Y_3^0 \mid G=0, Q=0). \qquad (\mathrm{ID}'_{TD})$$

Compared with the panel DD ID condition $E(\Delta Y_3^0 \mid Q=1) - E(\Delta Y_3^0 \mid Q=0) = 0$, $\mathrm{ID}'_{TD}$ (i.e., $\mathrm{ID}_{TD}$) allows this type of difference to be non-zero. Instead, $\mathrm{ID}'_{TD}$ requires the difference for $G=1$ to be the same as the difference for $G=0$. $\mathrm{ID}_{TD}$ is more general than $\mathrm{ID}_{DD}$ in this sense, which is, however, not exactly true because TD involves the extra grouping by $G$. As DD can be reduced to a BA if one BA is zero, TD can be reduced to a DD if one DD is zero.

### 6.2. Estimation for TD

Consider a panel linear model with the treatment applied from $\tau$ onwards:

$$Y_{it} = \beta_t + \beta_g G_i + \beta_q Q_i + \beta_{gq} G_i Q_i + \beta_d G_i Q_i 1[\tau \le t] + \delta_i + U_{it};$$

$W_{it}'\beta_w$ can be added to this model to control $W_{it}$. If desired, $\beta_g$ or $\beta_q$ can be allowed to be time-varying (i.e., $\beta_{gt}$ or $\beta_{qt}$) as the intercept $\beta_t$ is. But $\beta_{gq}$ cannot be time-varying (i.e., no $\beta_{gqt}$), because the treatment effect $\beta_d$ cannot be then separated from $\beta_{gqt}$.

First-differencing the $Y_{it}$ equation renders

$$\Delta Y_{it} = \Delta\beta_t + \beta_d G_i Q_i 1[t=\tau] + \Delta U_{it}.$$

$G_i$ or $Q_i$ may appear in the form $G_i\Delta\beta_{gt}$ or $Q_i\Delta\beta_{qt}$, if $\beta_{gt}$ or $\beta_{qt}$ is used. As in DD, $\beta_d$ can be time-varying, and $\tau$ can be individual-variant. Also, one-shot treatment $D_{it} = G_i Q_i 1[\tau = t]$ may happen, which makes, however, little

difference from the lasting treatment $G_iQ_i1[\tau \le t]$ as far as the estimation goes, once we allow for time-varying effects with $\sum_{a=0}^{T-\tau_i} \beta_{da} G_i Q_i 1[t = \tau_i + a]$. These extensions can be handled as in DD, by taking $GQ$ as a single qualification.

As an empirical example, in one store $(Q=1)$ of a national chain, Chetty et al. (2009) displayed the sales-tax-included price below the before-tax price for 3 categories of goods for 3 weeks (Feb. 22 to March 15, 2006) to see if showing the tax-inclusive price makes any difference in the patrons' purchase behavior. The three categories constitute the $G=1$ group, and two other stores in nearby cities constitute the $Q=0$ group. Since the tax has to be paid at the checkout point with the tax rate known, the tax-inclusive price is supposed to make no difference if the patrons are rational.

| Effect of Tax-Inclusive Price on Quantity Sold (SE): No W Controlled | | | |
|---|---|---|---|
| | Before | After | BA |
| $Q=1$ : treatment category ( $G=1$ ) | 25.17 | 23.87 | -1.30 |
| control category ( $G=0$ ) | 26.48 | 27.32 | 0.84 |
| DD for $Q=1$ | | | -2.14 (0.68) |
| $Q=0$ : treatment category ( $G=1$ ) | 27.94 | 28.19 | 0.25 |
| control category ( $G=0$ ) | 30.57 | 30.76 | 0.19 |
| DD for $Q=0$ | | | 0.06 (0.95) |
| TD | | | *-2.20 (0.59)* |

One of their estimation results are in Table 'Effect of Tax-Inclusive Price on Quantity Sold', which shows a negative effect of tax-inclusive price (surprise?). Since the DD for $Q=0$ in the bottom half little differs from zero, the DD for $Q=1$ would have been adequate with its effect −2.14 that differs little from the TD effect −2.20.

Turning to TD estimation with RCS, with $S_{it} \equiv 1[S_i = t]$, a RCS model for TD derived from the above panel linear model without $\delta_i$ is

$$Y_i = \beta_1 + \sum_{t=2}^{T} \Delta\beta_{t1} S_{it} + \beta_g G_i + \beta_q Q_i + \beta_{gq} G_i Q_i + \beta_d D_i + U_i$$

where $D_i = G_i Q_i 1[\tau \le S_i]$.

Do the OLS of $Y$ on $(1, S_2, \ldots, S_T, G, Q, GQ, D)$. Extensions of this model to many periods, time-varying treatment, individual-varying treatment timing, etc. can be handled as in DD.

### 6.3. Time-Wise Triple Difference (GDD) with Panel Data

If one more pre-treatment period is available, one extra difference in TD can be

done time-wise instead of cross-section group-wise, as this was the case with DD: DD with two cross-section group-wise BA's, and two time-wise BA's $BA_{23} - BA_{12}$. With $t = 1,2,3$ (2 pre-treatment periods), "time-wise TD" (Lee 2016b) is

$$DD_{23} - DD_{12}.$$

Since the effect $\beta_d$ of the DD treatment $Q[t = 3]$ still remains in $DD_{23}$, this time-wise TD can also identify $\beta_d$.

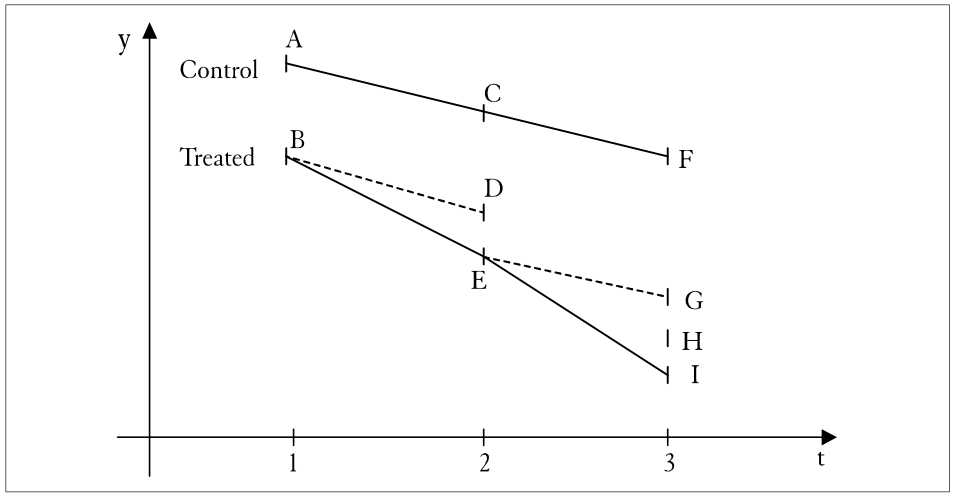
More generally, a time-wise QD is $QD_{123} \equiv TD_{23} - TD_{12}$, and another time-wise QD is

$$QD_{0123} \equiv DD_{23} - DD_{12} - (DD_{12} - DD_{01}) \text{ when } t = 0,1,2,3 \text{ are available.}$$

All of TD, time-wise TD and QD may be called 'generalized DD (GDD)', but to facilitate referencing, we use GDD only for time-wise TD $(DD_{23} - DD_{12})$ in the following.

To understand GDD better, examine Figure 'DD versus GDD'. DD uses only $t = 2,3$ to find the DD effect $\overline{GI}$, as DD regards $\overline{EG}$ as the untreated path of the T group because $\overline{EG}$ is parallel to $\overline{CF}$. Differently from this, GDD uses three periods 1, 2 and 3, and constructs the untreated path $\overline{EH}$ for the T group as a straight line extension of $\overline{BE}$; the GDD effect is $\overline{HI} = \overline{GI} - \overline{GH} = \overline{GI} - \overline{DE}$, which is nothing but $DD_{23} - DD_{12}$. That is, GDD takes into account the vertical difference $\overline{DE}$ that is $DD_{12}$, and subtracts $\overline{DE} = \overline{GH}$ from the $DD_{23}$ effect $\overline{GI}$ to come up with the GDD effect $\overline{HI}$.

**[Figure 2]** DD versus GDD



GDD has the ID condition:

$$E(\Delta Y_3^0 \mid Q=1)-E(\Delta Y_3^0 \mid Q=0)=E(\Delta Y_2^0 \mid Q=1)-E(\Delta Y_2^0 \mid Q=0). \qquad (\mathrm{ID}_{GDD})$$

The left-hand side being zero that is required in DD is relaxed; strictly speaking, however, this is not a relaxation because $\mathrm{ID}_{GDD}$ involves $t=1$ that does not appear in $\mathrm{ID}_{DD}$. To better see that $\mathrm{ID}_{GDD}$ relaxes $\mathrm{ID}_{DD}$, observe

$$Y_{it}^0=\beta_t+\beta_{q0}Q_i+\beta_{q1}tQ_i \Rightarrow \Delta Y_{it}^0=\Delta\beta_t+\beta_{q1}Q_i$$
$$\Rightarrow E(\Delta Y_t^0 \mid Q=1)-E(\Delta Y_t^0 \mid Q=0)=\beta_{q1}.$$

$\mathrm{ID}_{DD}$ is violated due to $\beta_{q1}\neq 0$, but GDD allows $\beta_{q1}\neq 0$ because both sides of $\mathrm{ID}_{GDD}$ equal $\beta_{q1}$. That is, GDD allows the $G$-group effect $\beta_{q0}Q+\beta_{q1}tQ$ to change over time due to $\beta_{q1}\neq 0$. Going further, $QD_{1234}$ allows $\beta_{q2}t^2Q$.

As for GDD estimation, we can allow non-parallel paths with $\beta_{q1}tQ_i$ in a panel model:

$$Y_{it}=\beta_t+\beta_{q0}Q_i+\beta_{q1}tQ_i+\beta_dQ_i1[t=3]+\delta_i+U_{it}.$$

Difference this model to get, at $t=3$ and $t=2$,

$$\Delta Y_{i3}=\beta_3-\beta_2+\beta_{q1}Q_i+\beta_dQ_i+\Delta U_{i3} \quad \text{and} \quad \Delta Y_{i2}=\beta_2-\beta_1+\beta_{q1}Q_i+\Delta U_{i2}$$
$$\Rightarrow \Delta Y_{i3}-\Delta Y_{i2}=(\beta_3-2\beta_2+\beta_1)+\beta_dQ_i+\Delta U_{i3}-\Delta U_{i2}.$$

The slope of $Q$ in the $\Delta Y_3$ equation is $\beta_{q1}+\beta_d$, which becomes $\beta_d$ only under the DD assumption $\beta_{q1}=0$. For GDD, $\beta_d$ is the slope of $Q$ for $\Delta Y_3-\Delta Y_2$ regardless of $\beta_{q1}=0$ or not. For QD, $\beta_d$ is the slope of $Q$ for $(\Delta Y_3-\Delta Y_2)-(\Delta Y_2-\Delta Y_1)$ regardless of $\beta_{q1}=0$ or $\beta_{q2}=0$ in $\beta_{q1}tQ+\beta_{q2}t^2Q$.

The differenced panel linear models can be estimated, which would be called 'fixed-effect estimation'. When $T>2$, instead of differencing, we may do 'within-group transformation' to remove $\delta_i$ by subtracting the temporally averaged model: with $\overline{Y}_{i\cdot}\equiv T^{-1}\sum_{s=1}^T Y_{is}$ and $\overline{U}_{i\cdot}\equiv T^{-1}\sum_{s=1}^T U_{is}$,

$$Y_{it}-\overline{Y}_{i\cdot}=\beta_t-\frac{1}{T}\sum_{s=1}^T\beta_s+\beta_{q1}Q_i\left(t-\frac{1}{T}\sum_{s=1}^T s\right)+\beta_dQ_i\left(1[\tau\le t]-\frac{1}{T}\sum_{s=1}^T 1[\tau\le s]\right)+U_{it}-\overline{U}_{i\cdot}$$
$$=\left(\beta_t-\frac{1}{T}\sum_{s=1}^T\beta_s\right)+\beta_{q1}\left\{t-\frac{T(T+1)}{2}\right\}Q_i+\beta_d\left\{1[\tau\le t]-\frac{T-\tau+1}{T}\right\}Q_i+U_{it}-\overline{U}_{i\cdot}$$

where both $\beta_{q1}$ and $\beta_d$ can be estimated, along with the intercept.

Instead of differencing or demeaning, $tQ$ (and $t^2Q$) can be added to the model to explicitly allow for non-parallel paths and the level model can be estimated as such, which would be called 'random-effect estimation'. The advantages/ disadvantages of these two approaches are analogous to those of random- versus fixed-effect approaches in panel data estimation: differencing has the advantage of alleviating endogeneity problems, but it loses time-constant regressors; the opposite holds for the "$tQ$-inserting" approach.

DD & GDD Effects of At-Home Service

| Parameter & Regressor | DD Estimate (tv) | GDD Estimate (tv) |
|---|---|---|
| $\beta_q, \beta_{0q}$ for $Q$ | -0.14 (-5.44) | -0.36 (-5.16) |
| $\beta_{1q}$ for $tQ$ | | 0.11 (3.41) |
| $\beta_d$ for $D$ | *0.09 (2.02)* | *-0.11 (-1.50)* |

As an empirical example for GDD, let $D$ be a new at-home service from 2013 onwards for the severely disabled $(Q=1)$ in South Korea, and $Y$ be their life satisfaction; those with $Q=0$ are disabled, but not severely. Table 'DD & GDD Effects of At-Home Service' reveals that DD based on ordered probit in Kim and Lee (2017) shows a significant effect $\beta_d = 0.09$, but GDD does not. The significantly positive $\beta_{1q} = 0.11$ in GDD indicates that the untreated responses did not move parallel across the two groups: the false positive DD effect 0.09 is due to omitting $tQ$.

Kim and Lee (2017) searched for the reason why $\beta_{1q} > 0$, and found that a new disability pension started before 2013 for the severely disabled. That is, two treatments took place during the period of interest, which hindered finding the effect of the second treatment $D$. The moral of this empirical example is "*apply DD to quiet periods with no other treatment than D*".

## VII. Time-Varying Qualification

So far we have been assuming that $Q$ is time-constant, which is, however, restrictive because $Q$ is often based on time-varying variables such as income, wealth, number of children, or residential location. When $Q$ is time-varying in DD, one may just try to use $D_{it} = Q_{it} 1[\tau \le t]$ as the treatment and proceed as usual, but this would be ignoring "*untreated moving effects*". that can confound the treatment effect. When $Q$ changes to make an individual newly (dis-) qualified and then (un-) treated, a change in $Y$ can be due to either the $Q$ change or the $D$ change, which is a confounding. Here, the $Y$ change due to the $Q$ change,

but not due to any $D$ change, is an untreated moving effect. This section examines how to uncover the desired treatment effect when untreated moving effects are present.

### 7.1. Untreated Moving Effect and Stayer Effect

Suppose that a minimum wage law goes into effect and the interest is on whether the law decreases work hours $Y_{it}$ or not. There are low-paying sectors such as fast-food industry affected by the law, and high-paying sectors such as financial industry not affected by the law. With $Q_{it}=1$ if person i works in a low-paying sector at period $t$, there are four groups based on $(Q_2, Q_3)$:

| $Q_3=0$: in high-paying at $t=3$ | $Q_3=1$: in low-paying at $t=3$ |
|---|---|
| $Q_2=0, Q_3=0$: "out-stayers" | $Q_2=0, Q_3=1$: "in-movers" |
| $Q_2=1, Q_3=0$: "out-movers" | $Q_2=1, Q_3=1$: "in-stayers" |

Suppose that the minimum wage law has no effect at all, with the in-stayers and out-stayers having $\Delta Y_3=0$. The in-movers lost a high-paying job and thus they work more ($E(\Delta Y_3 \mid \text{in-movers}) > 0$) to make up for the lost income, and the out-movers newly found a high-paying job to work less ($E(\Delta Y_3 \mid \text{out-movers}) < 0$). Hence,

| $E(\Delta Y_3 \mid Q_3=0)<0$ due to | $E(\Delta Y_3 \mid Q_3=1)>0$ due to |
|---|---|
| $E(\Delta Y_3 \mid Q_2=0, Q_3=0)=0$ | $E(\Delta Y_3 \mid Q_2=0, Q_3=1)>0$ |
| $E(\Delta Y_3 \mid Q_2=1, Q_3=0)<0$ | $E(\Delta Y_3 \mid Q_2=1, Q_3=1)=0$ |

This makes the conventional DD automatically positive, because

'conventional DD' $E(\Delta Y_3 \mid Q_3=1) - E(\Delta Y_3 \mid Q_3=0) > 0$.

Incidentally, Card and Krueger (1994) using fast food restaurant data reported positive effects on $Y_{it}$ of a minimum wage increase, where this kind of individual moves are not dealt with.

The problem of untreated moving effect matters more in RCS where $E(\Delta Y_3 \mid Q_3=1)$ equals $E(Y \mid Q=1, S=1) - E(Y \mid Q=1, S=0)$, because the available information on $Q$ would be only on $Q_3$, not on $Q_2$. If panel data are available, however, Lee and Kim (2014) showed that the problem can be avoided using

$$\text{'(panel) stayer DD'}: \; E(\Delta Y_3 \mid Q_2=1, Q_3=1) - E(\Delta Y_3 \mid Q_2=0, Q_3=0)$$
$$= E(Y_3^1 - Y_2^0 \mid Q_2=1, Q_3=1) - E(Y_3^0 - Y_2^0 \mid Q_2=0, Q_3=0). \quad (7.1)$$

The ID condition for stayer DD is

$$E(\Delta Y_3^0 \mid Q_2 = 1, Q_3 = 1) = E(\Delta Y_3^0 \mid Q_2 = 0, Q_3 = 0). \qquad (\mathrm{ID}_{SDD})$$

Under this, *stayer DD becomes the effect on the in-stayers (not on* $Q_3 = 1$*) at the post-treatment period* $t = 3$:

$$E(Y_3^1 - Y_3^0 \mid Q_2 = 1, Q_3 = 1).$$

One might think that the untreated moving effect problem can be avoided by changing the observation unit from an individual to a fixed establishment such as shop or region, but this does not solve the problem because the composition of shops or regions change as individuals move around. That is, if a fixed establishment is the observation unit, then its composition should be controlled in finding the desired treatment effect.

### 7.2. Panel Linear Model with Source-Dependent Effect

With $V_{it} = \delta_i + U_{it}$, a panel linear model is

$$Y_{it}^0 = \beta_t + \beta_q Q_{it} + V_{it} \quad \text{and} \quad Y_{it}^1 = Y_{it}^0 + \beta_s Q_{i,t-1} + \beta_m (1 - Q_{i,t-1}).$$

When treated at $t$, the intercept shifts by $\beta_s$ if $Q_{t-1} = 1$, or by $\beta_m$ if $Q_{t-1} = 0$; $s$ in $\beta_s$ for 'stayers', and $m$ in $\beta_m$ for 'movers'. The treatment effect depends on the source $Q_{t-1}$, and we thus have a "source-dependent effect". More generally, the effect may depend on the "path" $Q_{t-1}$, $Q_{t-2}$, $Q_{t-3}$ and so on. In the $Y_{it}^0$ equation, $\beta_q Q_{it}$ represents an untreated moving effect.

With $D_{it} = Q_{it} 1[t = 3]$, the observed response is

$$\begin{aligned} Y_{it} &= \beta_t + \{\beta_s Q_{i,t-1} + \beta_m (1 - Q_{i,t-1})\} \cdot Q_{it} 1[t = 3] + \beta_q Q_{it} + V_{it} \\ &= \beta_t + \beta_s 1[t = 3] Q_{i,t-1} Q_{it} + \beta_m 1[t = 3] (1 - Q_{i,t-1}) Q_{it} + \beta_q Q_{it} + V_{it} \end{aligned}$$

where $\{\cdot\}$ is the treatment effect. First-differencing the $Y_{it}$ model yields, for $t = 3$,

$$\Delta Y_{i3} = \Delta\beta_3 + \beta_s Q_{i2} Q_{i3} + \beta_m (1 - Q_{i2}) Q_{i3} + \beta_q \Delta Q_{i3} + \Delta U_{i3}.$$

OLS/IVE can be applied to this $\Delta Y_3$ model where the regressors are

$$\Lambda_i \equiv \{1, Q_{i2}Q_{i3}, (1-Q_{i2})Q_{i3}, \Delta Q_{i3}\} \quad \text{for} \quad \gamma \equiv (\Delta\beta_3, \beta_s, \beta_m, \beta_q)' .$$

Two periods give 4 groups based on $Q_2 = 0,1$ and $Q_3 = 0,1$, with which the 4 parameters in $\gamma$ are identified; the model is fully saturated in this sense. The slope $\beta_q$ of $\Delta Q_3$ is the untreated moving effect; if $Q$ were time-constant, $\beta_q Q$ would have dropped out of $\Delta Y_3$. Two treatment effects, $\beta_s$ and $\beta_m$, appear in the model for the two treated groups $(Q_2 = 0, Q_3 = 1)$ and $(Q_2 = 1, Q_3 = 1)$; both have $Q_3 = 1$.

In the above panel "source-dependent-effect" model, assuming $E(Q_2 \mid Q_3) = E(Q_2)$ and $E(\Delta U_3 \mid Q_3) = 0$, we have

$$E(\Delta Y_3 \mid Q_3 = 1) = \Delta\beta_3 + \beta_s E(Q_2) + \beta_m \{1 - E(Q_2)\} + \beta_q \{1 - E(Q_2)\},$$
$$E(\Delta Y_3 \mid Q_3 = 0) = \Delta\beta_3 + \beta_q \{0 - E(Q_2)\}.$$

From this, we can see that the conventional DD is contaminated by $\beta_q$ because

$$E(\Delta Y_3 \mid Q_3 = 1) - E(\Delta Y_3 \mid Q_3 = 0) = \beta_s E(Q_2) + \beta_m \{1 - E(Q_2)\} + \beta_q .$$

Even if $\beta_s = \beta_m$, still $\beta_q$ remains because this becomes $\beta_s + \beta_q$.

As an empirical example, in January 2008, South Korea started the 'Basic Elder Pension (BEP)' for persons of age $\geq 65$. Lee and Kim (2014) examined the effect of BEP on $Y = \ln(\text{health care expenditure})$, where $Q$ depends on income/wealth being lower than a cutoff. Table 'Conventional Panel DD' shows that the conventional DD gives an 18% BEP effect along with the untreated moving effect of −15%:

| Conventional Panel DD: OLS (tv) for $\Delta Y_t$ : $N = 2046$ | |
|---|---|
| $Q_3$ (treatment) | 0.182 (2.44) |
| $\Delta Q_3$ (untreated moving: getting poorer) | -0.150 (-2.11) |

In contrast, using the above source-dependent-effect model, Table 'Stayer Panel DD' shows $\beta_s \neq \beta_m$ and an insignificant untreated moving effect.

| Stayer Panel DD: OLS (tv) for $\Delta Y_t$ : $N = 2046$ | |
|---|---|
| $Q_2 Q_3$ (in-stayer: getting BEP without getting poorer) | 0.162 (2.10) |
| $(1-Q_2)Q_3$ (in-mover: getting BEP while getting poorer) | 0.025 (0.14) |
| $\Delta Q_3$ (untreated moving: no BEP while getting poorer) | -0.088 (-0.91) |

### 7.3. Overcoming Ashenfelter Dip Problem with Stayers

The well-known 'Ashenfelter (1978, p.51) dip' for job trainings is that the T group experience a dip (i.e., a low $Y_2$ in earnings) just before getting treated: "*parts of the observed earnings increase following training may merely be a return to a permanent path of earnings that was temporarily interrupted, ..., considerable ambiguity in untangling the effect of training from the effect of this transitory phenomenon*." Since the 'dip' is transitory by definition, the T group is bound to have a higher post-treatment earnings $Y_3$ even without the treatment—an untreated moving effect of a sort. Stayer DD can take care of this problem as follows.

Suppose a training is given to the unemployed: $Q_t = 1[Y_{t-1} = 0]$. There are the persistently unemployed $(Y_1 = 0, Y_2 = 0) \Leftrightarrow (Q_2 = 1, Q_3 = 1)$, and the temporarily unemployed $(Y_1 > 0, Y_2 = 0) \Leftrightarrow (Q_2 = 0, Q_3 = 1)$. The stayer DD overcomes the Ashenfelter dip problem, as the movers are either not used in the stayer DD (7.1), or the three effects ($\beta_s$, $\beta_m$ and $\beta_q$) are separately identified in the source-dependent-effect model.

Essentially the same problem occurs in evaluating an education program. Suppose the program is applied to a low-score group $Q = 1$ that consists of permanently low-score students and temporarily low-score students. In the subsequent test, some of the temporarily low-score students are bound to bounce back, which can give a false impression that the program is effective. Using only stayers can avoid biases like this.

## References

Abadie, A. (2005), "Semiparametric Difference-in-differences Estimators," *Review of Economic Studies*, 72, 1–19.

Abadie, A., A. Diamond and J. Hainmueller (2010), Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505.

_____ (2015), "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59, 495–510.

Abadie, A., and J. Gardeazabal (2003), "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 112–132.

Ando, M. (2015), "Dreams of Urbanization: Quantitative Case Studies on the Local Impacts of Nuclear Power Facilities Using the Synthetic Control Method," *Journal of Urban Economics*, 85, 68–85.

Angrist, J. D. and A. B. Krueger (1999), "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, 3A, edited by O. Ashenfelter and D. Card, North-Holland.

Angrist, J. D. and J. S. Pischke (2009), *Mostly Harmless Econometrics*, Princeton University Press.

Ashenfelter, O. (1978), "Estimating the Effect of Training Program on Earnings," *Review of Economics and Statistics*, 60, 47–57.

Athey, S. and G. W. Imbens (2006), Identification and Inference in Nonlinear Difference-in-differences Models," *Econometrica*, 74, 431–497.

Bertrand, M., E. Duflo and S. Mullainathan (2004), "How much Should we Trust Differences-in-differences Estimates," *Quarterly Journal of Economics*, 119, 249–275.

Besley, T. and A. Case (2000), "Unnatural Experiments? Estimating the Incidence of Endogenous Policies," *Economic Journal*, 110, F672–F694.

Bohn, S., M. Lofstrom and S. Raphael (2014), "Did the 2007 Legal Arizona Workers Act Reduce the State's Unauthorized Immigrant Population?" *Review of Economics and Statistics*, 96, 258–269.

Brewer, M., T. F. Crossley and R. Joyce (2018), Inference with Difference-in-differences Revisited," *Journal of Econometric Methods*, 7, doi: 10.1515/jem-2017-0005.

Campbell, K. B. and C. Brakewood (2017), "Sharing Riders: How Bikesharing Impacts Bus Ridership in New York City," *Transportation Research Part A*, 100, 264–282.

Card, D. (1990), "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review*, 43, 245–257.

Card, D. and A. B. Krueger (1994), "Minimum Wage and Employment: A Case Study of the Fast-food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772–793.

Chemin M. and E. Wasmer (2009), "Using Alsace-Moselle Local Laws to Build a Difference-in-differences Estimation Strategy of the Employment Effects of the 35-hour Workweek Regulation in France," *Journal of Labor Economics*, 27, 487–524.

Chetty, R. A. Looney and K. Kroft (2009), Salience and Taxation: Theory and Evidence," *American Economic Review*, 99, 1145–1177.

Choi, S., A. Pellen and V. Masson (2017), How Does Daylight Saving Time Affect Electricity Demand? An Answer Using Aggregate Data from a Natural Experiment in Western Australia," *Energy Economics*, 66, 247–260.

Chung, Y. W., S. J. Kang, T. Matsubayashi, Y. Sawada and M. Ueda (2016), "The Effectiveness of Platform Screen Doors for the Prevention of Subway Suicides in South Korea," *Journal of Affective Disorders*, 194, 80–83.

De Chaisemartin, C. and X. D'Haultfoeuille (2018), "Fuzzy Difference in Differences," *Review of Economic Studies*, 85, 999–1028.

Doudchenko, N. and G. W. Imbens (2016), "Balancing, Regression, Difference-in-differences and Synthetic Control Methods," NBER working paper 22791.

DuPont W., I. Noy, Y. Okuyama and Y. Sawada (2015), "The Long-run Socio-economic Consequences of a Large Disaster: The 1995 Earthquake in Kobe," PLoS ONE 10(10): e0138714.

Eissa, N. and J. B. Liebman (1996), "Labor Supply Response to the Earned Income Tax Credit," *Quarterly Journal of Economics*, 111, 605–637.

Hagiwara, A., M. Ueyama, A. Ramlawi and Y. Sawada (2013), "Is the Maternal and Child Health (MCH) Handbook Effective in Improving Health-related Behavior?" Evidence from Palestine, *Journal of Public Health Policy*, 34, 31–45.

Heckman, J. J., R. J. Lalonde and J. A. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics*, 3B, edited by O.C. Ashenfelter and D. Card, North-Holland.

Helland, E. and A. Tabarrok (2007), "Does Three Strikes Deter? A Nonparametric Estimation," *Journal of Human Resources*, 22, 309–330.

Hsieh, C. T., S. Shimizutani and M. Hori (2010), "Did Japan's Shopping Coupon Program Increase Spending?" *Journal of Public Economics*, 94, 523–529.

Hwang, H. and M. J. Lee (2018), "A Simple Makeover Can Increase Bus Ridership," unpublished paper.

Kan, K. and M. J. Lee (2018), "The Effects of Education on Fertility: Evidence from Taiwan," *Economic Inquiry*, 56, 343–357.

Kan, K., S. K. Peng and P. Wang (2017), Understanding Consumption Behavior: Evidence from Consumers' Reaction to Shopping Vouchers," *American Economic Journal: Economic Policy*, 9, 137–153.

Kim, H. A., Y. S. Kim and M. J. Lee (2012), "Treatment Effect Analysis of Early Reemployment Bonus Program: Panel MLE and Mode-based Semiparametric Estimator for Interval Truncation," *Portuguese Economic Journal*, 11, 189–209.

Kim, K. M. and M. J. Lee (2019), "Difference in Differences in Reverse," *Empirical Economics*, 57, 705–725.

Kim, Y. S. and M. J. Lee (2017), "Ordinal Response Generalized Difference-in-differences with Varying Categories: The Health Effect of a Disability Program in Korea," *Health Economics*, 26, 1121–1131.

Kotchen, M. J. and L. E. Grant (2011), "Does Daylight Saving Time Save Energy?

Evidence from a Natural Experiment in Indiana," *Review of Economics and Statistics*, 93, 1172–1185.

Kreif, N., R. Grieve, D. Hangartner, A. J. Turner, S. Nikolova and M. Sutton (2016), "Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units," *Health Economics*, 25, 1514–1528.

Kuo, T. C. (2012), "Evaluating California Under-age Drunk Driving Laws: Endogenous Policy Lags," *Journal of Applied Econometrics*, 27, 1100–1115.

Lee, M. J. (2005), *Micro-econometrics for Policy, Program, and Treatment Effects*, Oxford University Press.

_____ (2016a), *Matching, Regression Discontinuity, Difference in Differences, and Beyond*, Oxford University Press.

_____ (2016b), "Generalized Difference in Differences with Panel Data and Least Squares Estimator," *Sociological Methods & Research*, 45, 134–157.

_____ (2018), "Simple Least Squares Estimator for Treatment Effects Using Propensity Score Residuals," *Biometrika*, 105, 149–164.

_____ (2019), "How to set up Outcome Model Using Instrument Score to Minimize Treatment Confounding Problem," unpublished paper.

Lee, M. J. and C. H. Kang (2006), "Identification for Difference in Differences with Cross-section and Panel Data," *Economics Letters*, 92, 270–276.

Lee, M. J. and Y. S. Kim (2014), "Difference in Differences for Stayers with a Time-varying Qualification: Health Expenditure Elasticity of the Elderly," *Health Economics*, 23, 1134–1145.

Mahmud M. and Y. Sawada (2018), "Infrastructure and Well-being: Employment Effects of Jamuna Bridge in Bangladesh," *Journal of Development Effectiveness*, 10, 327–340.

Monstad, K., C. Propper and K. G. Salvanes (2008), Education and Fertility: Evidence from a Natural Experiment," *Scandinavian Journal of Economics*, 110, 827–852.

Munasib, A. and D. S. Rickman (2015), Regional Economic Impacts of the Shale Gas and Tight Oil Boom: A Synthetic Control Analysis," *Regional Science and Urban Economics*, 50, 1–17.

Puhani, P. A. (2012), The Treatment Effect, the Cross Difference, and the Interaction Term in Nonlinear "Difference in Differences" Models," *Economics Letters*, 115, 85–87.

Xu, Y. (2017), "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, 25, 57–76.