CHAPTER 12

Program Evaluation Using Panel Data

12.1 INTRODUCTION

Individuals are often given “treatments,” such as a drug trial or a training program. Let y_{it}^{1*} and y_{it}^{0*} denote the outcome of receiving the treatment and not receiving the treatment for the ith individual at time t; then the treatment effect is simply

$$\Delta_{it} = y_{it}^{1*} - y_{it}^{0*}, \qquad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T \end{aligned} \tag{12.1.1}$$

The *average treatment effect* (ATE)[1] at time t is

$$\text{ATE}(t) = E(\Delta_{it}) = \Delta_t. \tag{12.1.2}$$

The ATE across individuals and over time is

$$\text{ATE} = E(\Delta_t) = E(E(\Delta_{it})) = \Delta. \tag{12.1.3}$$

Let y_{it}^1 and y_{it}^0 denote the realized y_{it}^{1*} and y_{it}^{0*}, respectively. If both y_{it}^1 and y_{it}^0 are observed, then the measurement of Δ_{it}, Δ_t, or Δ are straightforward. However, in most cases the observed data take the form

$$y_{it} = d_{it} y_{it}^1 + (1 - d_{it}) y_{it}^0, \tag{12.1.4}$$

where d_{it} is the treatment status dummy that takes the value 1 if the ith unit at t receives the treatment and 0 if not.

When either y_{it}^1 or y_{it}^0 are missing, the estimation of treatment effects involves the construction or prediction of missing or never realized y_{it}^{1*} or y_{it}^{0*}. We call those units with $y_{it} = y_{it}^1$ the *treatment group*, and those units with $y_{it} = y_{it}^0$ the *control group*. Then the construction of counterfactuals for the missing y_{it}^{1*} or y_{it}^{0*} has to rely on the information from the treatment group or the control group. However, the observed data could be subject to *selection on observables* and/or *selection on unobservables*.

By *selection on observables*, we suppose the observed outcomes are the sum of observable factors, $\boldsymbol{x}_{it}$, and unobserved factors,

$$y_{it}^1 = g_1(\boldsymbol{x}_{it}) + \varepsilon_{it}^1, \tag{12.1.5}$$

$$y_{it}^0 = g_0(\boldsymbol{x}_{it}) + \varepsilon_{it}^0, \tag{12.1.6}$$

[1] See Heckman (1997), Heckman and Vytlacil (2001), and Imbens and Angrist (1994) for the definitions of the marginal treatment effect (MTE) and the local average treatment effect (LATE).

where the unobservable factors ε_{it}^1 and ε_{it}^0 satisfy $E(\varepsilon_{it}^1|\boldsymbol{x}_{it}) = E(\varepsilon_{it}^0|\boldsymbol{x}_{it}) = 0$. Then the ATE at a given $\boldsymbol{x}_{it}$ is

$$\begin{aligned}\text{ATE}(\boldsymbol{x}_{it}) &= g_1(\boldsymbol{x}_{it}) - g_0(\boldsymbol{x}_{it}) + E(\varepsilon_{it}^1|\boldsymbol{x}_{it}) - E(\varepsilon_{it}^0|\boldsymbol{x}_{it}) \\ &= g_1(\boldsymbol{x}_{it}) - g_0(\boldsymbol{x}_{it}).\end{aligned} \tag{12.1.7}$$

However, if y_{it}^0 is missing, the construction of $g_0(\boldsymbol{x}_{it})$ has to rely on the information from control groups in the observed data. The treatment group and control group could be drawn from different populations (e.g., Dehejia and Wahba 1999; Lalonde 1986), and those individuals in the control group may have different values of $\boldsymbol{x}_{it}$ from those in the treatment group.

By *selection on unobservables*, we mean the participation decision d_{it} could be correlated with potential outcomes y_{it}^{1*} or y_{it}^{0*}. Then, although $E(\varepsilon_{it}^1|\boldsymbol{x}_{it}) = E(\varepsilon_{it}^0|\boldsymbol{x}_{it}) = 0$,

$$E(\varepsilon_{it}^1|\boldsymbol{x}_{it}, d_{it}) \neq 0. \tag{12.1.8}$$

and/or

$$E(\varepsilon_{it}^0|\boldsymbol{x}_{it}, d_{it}) \neq 0. \tag{12.1.9}$$

In other words, the observed sample could be subject to sample selection effect (Heckman 1979, Chapter 7).

In Section 12.2 we review the parametric, semiparametric, and nonparametric approach to correcting the *selection on observables* or *unobservables effects* with *cross-sectional data*. Section 12.3 discusses the panel nonparametric approach, semiparametric approach, and parametric approach, respectively. Section 12.4 presents two applied examples, measuring the impact of closer economic partnerships between Hong Kong and Mainland China on the Hong Kong economy, and the impact of the California tobacco control program on California's per capita cigarette consumption and health expenditure. Section 12.5 considers issues arising with multiple treated units. Section 12.6 discusses simulating and ranking of program impacts under different policy options.

12.2 CROSS-SECTIONAL DATA APPROACH

When only cross-sectional data are available, $T = 1$. For notational simplicity, we shall drop the subscript t for a variable, namely, we shall say $y_{it} = y_i$. The fundamental assumption for modeling cross-sectional data is that $E(y_i|\boldsymbol{x}_i = \boldsymbol{a}) = E(y_j|\boldsymbol{x}_j = \boldsymbol{a})$, i.e., conditional on $\boldsymbol{x}_i = \boldsymbol{x}_j = \boldsymbol{a}$, only ATE($\boldsymbol{x}$) or a subgroup of ATE can be estimated, say, the *average treatment effects of the treated* (ATET),

$$\text{ATET}(\boldsymbol{x}) = E\left[(y_i^{1*}|\boldsymbol{x}_i = \boldsymbol{x}, d_i = 1) - (y_i^{0*}|\boldsymbol{x}_i = \boldsymbol{x}, d_i = 1)\right]. \tag{12.2.1}$$

or ATE of *the untreated* (ATEU)

$$\text{ATEU}(\boldsymbol{x}) = E\left[(y_i^{1*}|\boldsymbol{x}_i = \boldsymbol{x}, d_i = 0) - E(y_i^{0*}|\boldsymbol{x}_i = \boldsymbol{x}, d_i = 0)\right]. \tag{12.2.2}$$

12.2.1 Parametric Approach

Suppose $y_i^1 = y_i^{1*}$ and $y_i^0 = y_i^{0*}$, (Equations 12.1.5 and 12.1.6), could be specified parametrically. Let the participation of treatment $d_i = 1$ if $d_i^* > 0$ and $d_i = 0$ if $d_i^* \leq 0$, where

$$d_i^* = h(z_i) + v_i, \tag{12.2.3}$$

and z_i denote the observed factors determining the selection equation that may overlap with some or all elements of $\boldsymbol{x}$. With a known joint distribution of $f(\varepsilon^1, \varepsilon^0, v)$, the mean functions $g_1(\boldsymbol{x}), g_0(\boldsymbol{x})$ can be consistently estimated by the maximum likelihood method and

$$\text{ATE}(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_0(\boldsymbol{x}), \tag{12.2.4}$$

e.g., Chapter 7 and Damrongplasit, Hsiao, and Zhao (2010).

12.2.2 Semiparametric Approach

Suppose $g_1(\boldsymbol{x})$ and $g_0(\boldsymbol{x})$ are well specified, but not $h(\cdot)$ or $f(\varepsilon^1, \varepsilon^0, v)$. Then $g_1(\boldsymbol{x})$ and $g_0(\boldsymbol{x})$ could still be estimated by the semiparametric method discussed in Section 7.1, say, the Robinson (1988a) partial linear regression method or the Ahn and Powell (1993) pairwise difference method. However, some of the coefficients of the variables that appear in $g_1(\cdot), g_0(\cdot)$, and $h(\cdot)$ may not be consistently estimated; say the intercepts, which could be crucial in estimating the treatment effects (e.g., Liu et al. 2009).

12.2.3 Nonparametric Approach

If $g_1(\boldsymbol{x})$ and $g_0(\boldsymbol{x})$ are unspecified, they can be estimated by nonparametric methods provided that conditioning on a set of confounding variables, say $\boldsymbol{x}$, the distribution of (y^{1*}, y^{0*}) are independent of d (or d^*). In other words, conditional on $\boldsymbol{x}$, there is no selection on unobservables (*conditional independence or unconfoundness*),

$$E(y^{1*} \mid d, \boldsymbol{x}) = E(y^{1*} \mid \boldsymbol{x}), \tag{12.2.5}$$

$$E(y^{0*} \mid d, \boldsymbol{x}) = E(y^{0*} \mid \boldsymbol{x}). \tag{12.2.6}$$

Then, conditional on $\boldsymbol{x}$, the average treatment effect, ATE$(\boldsymbol{x})$,

$$\begin{aligned} \text{ATE}(\boldsymbol{x}) &= E(y^{1*} - y^{0*} \mid \boldsymbol{x}) \\ &= E(y \mid d = 1, \boldsymbol{x}) - E(y \mid d = 0, \boldsymbol{x}) \\ &= E(y^{1*} \mid \boldsymbol{x}) - E(y^{0*} \mid \boldsymbol{x}) \\ &= g_1(\boldsymbol{x}) - g_0(\boldsymbol{x}). \end{aligned} \tag{12.2.7}$$

However, the dimension of $\boldsymbol{x}$ could be large. For $\boldsymbol{x} = \boldsymbol{a}$, let $\psi = \{i | \boldsymbol{x}_i = a\}$, $n_a = \sum_{i \in \psi} 1\{\boldsymbol{x}_i = \boldsymbol{a}\}$ could be small, where $1(A) = 1$ if A occurs and zero otherwise. With a small number of observations, $E(y^{1*} | \boldsymbol{x})$ and $E(y^{0*} | \boldsymbol{x})$ could not be well estimated using the observed treatment and control group data. To overcome the curse of dimensionality, several methods have been proposed.

12.2.3.1 Matching Observables in Terms of Propensity Score Method (or Selection on Observables Adjustment)

Rosenbaum and Rubin (1983, 1985) have suggested a propensity score method to match the observable characteristics of the treatment group and the control group. The Rosenbaum and Rubin (1983) propensity score methodology supposes unit i has observable characteristics $\boldsymbol{x}_i$. Let $P(\boldsymbol{x}_i)$ be the probability of unit i having been assigned to treatment, called the *propensity score* in statistics and *choice probability* in econometrics, defined as

$P(\boldsymbol{x}_i) = \text{Prob}\ (d_i = 1 \mid \boldsymbol{x}_i) = E(d_i \mid \boldsymbol{x}_i)$. Assume that $0 < P(\boldsymbol{x}_i) < 1$ for all $\boldsymbol{x}_i$,[2] and Prob $(d_1, \ldots, d_N \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \prod_{i=1}^{N} P(\boldsymbol{x}_i)^{d_i}[1 - P(\boldsymbol{x}_i)]^{1-d_i}$, for $i = 1, \ldots, N$. If the treatment assignment is ignorable given $\boldsymbol{x}$, then it is ignorable given $P(\boldsymbol{x})$; that is,

$$\left\{(y_i^{1*}, y_i^{0*}) \perp d_i \mid \boldsymbol{x}_i\right\} \Longrightarrow \left\{(y_i^{1*}, y_i^{0*}) \perp d_i \mid P(\boldsymbol{x}_i)\right\}, \tag{12.2.8}$$

where $\perp$ denotes orthogonality.

To show that (12.2.8) holds, it is sufficient to show that

$$\text{Prob}\ (d = 1 \mid y^{0*}, y^{1*}, P(\boldsymbol{x})) = \text{Prob}\ (d = 1 \mid P(\boldsymbol{x})). \tag{12.2.9}$$

We note that

$$P(\boldsymbol{x}) = \ \text{Prob}\ (d = 1 \mid \boldsymbol{x}) = \ \text{Prob}\ (d = 1 \mid y^{0*}, y^{1*}, \boldsymbol{x}). \tag{12.2.10}$$

Equation (12.2.9) follows from applying the treatment assignment assumption to

$$\begin{aligned}
&\text{Prob}\ (d = 1 \mid y^{0*}, y^{1*}, P(\boldsymbol{x})) \\
&\qquad = E_x \left\{\text{Prob}\ (d = 1 \mid y_0^*, y_1^*, \boldsymbol{x}) \mid y^{0*}, y^{1*}, P(\boldsymbol{x})\right\} \\
&\qquad = E_x \left\{\text{Prob}\ (d = 1 \mid \boldsymbol{x}) \mid y^{0*}, y^{1*}, P(\boldsymbol{x})\right\} \\
&\qquad = E_x \left\{P(\boldsymbol{x}) \mid y^{0*}, y^{1*}, P(\boldsymbol{x})\right\} \\
&\qquad = E_x \{P(\boldsymbol{x}) \mid P(\boldsymbol{x})\} = P(\boldsymbol{x}),
\end{aligned} \tag{12.2.11}$$

where E_x denotes taking the expectation with respect to $\boldsymbol{x}$. Moreover, (12.2.8) also implies that

$$\boldsymbol{x}_i \perp d_i \mid P(\boldsymbol{x}_i). \tag{12.2.12}$$

To prove (12.2.12), it is sufficient to show that

$$\text{Prob}\ (d = 1 \mid \boldsymbol{x}, P(\boldsymbol{x})) = \ \text{Prob}\ (d = 1 \mid P(\boldsymbol{x})). \tag{12.2.13}$$

Equation (12.2.13) follows from Prob $(d = 1 \mid \boldsymbol{x}, P(\boldsymbol{x})) =$ Prob $(d = 1 \mid \boldsymbol{x}) = P(\boldsymbol{x})$ and

$$\begin{aligned}
\text{Prob}\ (d = 1 \mid P(\boldsymbol{x})) &= E_x\ \{\text{Prob}\ (d = 1 \mid \boldsymbol{x}, P(\boldsymbol{x})) \mid P(\boldsymbol{x})\} \\
&= E_x\ \{P(\boldsymbol{x}) \mid P(\boldsymbol{x})\} = P(\boldsymbol{x}).
\end{aligned}$$

Equation (12.2.12) implies that the conditional density of $\boldsymbol{x}$, given d and $P(\boldsymbol{x})$ is

$$f(\boldsymbol{x} \mid d = 1, P(\boldsymbol{x})) = f(\boldsymbol{x} \mid d = 0, P(\boldsymbol{x})) = f(\boldsymbol{x} \mid P(\boldsymbol{x})). \tag{12.2.14}$$

In other words, Equation (12.2.12) implies that if a subclass of units or a matched treatment-control pair is homogeneous in $P(\boldsymbol{x})$, then the treated and control units in that subclass or matched pair will have the same distribution of $\boldsymbol{x}$. It implies that, at any value of a propensity score, the mean difference between the treatment group and control group is an unbiased estimate of the average treatment effect at that value of the propensity score if treatment assignment is ignorable. The ATE is

[2] The assumption that $0 < P(\boldsymbol{x}_i) < 1$ guarantees that for each $\boldsymbol{x}_i$, we obtain observations in both the treated and untreated states. This assumption can be relaxed as long as there are $\boldsymbol{x}$ such that $0 < P(\boldsymbol{x}) < 1$.

$$\Delta(P(\boldsymbol{x})) = E\{E(y \mid d = 1, P(\boldsymbol{x})) - E(y \mid d = 0, P(\boldsymbol{x}))\}, \qquad (12.2.15)$$

where the outer expectation is over the distribution of $\{P(\boldsymbol{x}) \mid \boldsymbol{x}\}$.

The attraction of the propensity score matching method is that in (12.2.7) we condition on $\boldsymbol{x}$ (intuitively, to find observations with similar covariates), whereas in (12.2.15) we condition just on the propensity score because (12.2.15) implies that observations with the same propensity score have the same distribution of the full vector of covariates, $\boldsymbol{x}$. Equation (12.2.12) asserts that conditional on $P(\boldsymbol{x})$, the distribution of covariates should be the same across the treatment and comparison groups. In other words, conditional on the propensity score, each individual has the same probability of assignment to treatment, as in a randomized experiment. Therefore, the estimation of average treatment effect can be done in two steps. The first step involves the estimation of propensity score parametrically or nonparametrically (e.g., see Chapter 6). In the second step, given the estimated propensity score, one can estimate $E\{y \mid P(\boldsymbol{x}), d = j\}$ for $j = 0, 1$, and take the difference between the treatment and control groups, then weight these by the frequency of treated observations or frequency of (both treated and untreated) observations in each stratum to get an estimate of ATE (or ATET if the approximation is restricted to those in the treatment group, $d_i = 1$) $(E\{E[y \mid d = 1, P(\boldsymbol{x})] - E[y \mid d = 0, P(\boldsymbol{x})]\} = E\{E[y_1 - y_0 \mid P(\boldsymbol{x})]\})$, where the outer expectation is with respect to the propensity score, $P(\boldsymbol{x})$. For examples of using this methodology to evaluate the effects of training programs in nonexperimental studies, see Dehejia and Wahba (1999), Lalonde (1986), and Liu, Hsiao, Matsumoto, and Chou (2009). However, even the nonconfoundness (or no sample selection effects; Heckman 1979) holds, the computation of treatment effects can be sensitive to the way the breakup of a continous variable $P(\boldsymbol{x})$ into subintervals, say $P(\boldsymbol{x}) = a$ for any $a - c < P(\boldsymbol{x}) < a + c$, in order to compute estimated $E(y^{1*}|P(\boldsymbol{x}) = a)$ and $E(y^{0*}|P(\boldsymbol{x}) = a)$ from an observed sample (e.g., Damrongplasit et al. 2010).

12.2.3.2 Regression Discontinuity (RD) Design

Let $\boldsymbol{x}_i = (w_i, \boldsymbol{q}_i')$ be k covariates, where w_i is a scalar and $\boldsymbol{q}_i$ is a $(k-1) \times 1$ vector. Both w_i and $\boldsymbol{q}_i$ are not affected by the treatment. The basic idea behind the RD design is that assignment to the treatment is determined, either completely or partly, by the value of a predictor w_i being on either side of a fixed threshold. This predictor, w_i (together with $\boldsymbol{q}_i$), also affects the potential outcomes.

For notational ease, we shall assume $\boldsymbol{q}_i = 0$ for this subsection. In the sharp RD (SRD) designs, it is assumed that all units with the values of w at least c are assigned to the treatment group and participation is mandatory for these individuals, and with values of w less than c are assigned to the control groups and members of these group are not eligible for the treatment. Then,

$$\begin{aligned} \text{ATE}(c) &= \lim\nolimits_{w \downarrow c} E(y \mid w) - \lim\nolimits_{w \uparrow c} E(y \mid w), \\ &= E(y^1 - y^0 \mid w = c) \end{aligned} \qquad (12.2.16)$$

This approach assumes the unconfoundedness of Rosenbaum and Rubin (1983). However, it violates $0 < P(d = 1 \mid \boldsymbol{x}) < 1$. It assumes $P(d = 1|w > c) = 1$ and $P(d = 0|w < c) = 0$.

This approach assumes either: (i) $E(y^0 \mid w)$ and $E(y^1 \mid w)$ are continuous in w; or (ii) $F_{y^0|w}(y \mid w)$ and $F_{y^1|w}(y \mid w)$ are continuous in w for all y.

The Fuzzy RD (FRD) allows $0 < P(d = 1|\boldsymbol{x}) < 1$ but assumes

$$\lim\nolimits_{w \downarrow c} \text{Prob}(d = 1 \mid w) \neq \lim\nolimits_{w \uparrow c} \text{Prob}(d = 1 \mid w), \qquad (12.2.17)$$

then

$$\text{ATE}(c) = \frac{\lim_{w\downarrow c} E(y \mid w) - \lim_{w\uparrow c} E(y \mid w)}{\lim_{w\downarrow c} P(d=1 \mid w) - \lim_{w\uparrow c} P(d=1 \mid w)}. \tag{12.2.18}$$

To see this, let

$$\lim_{w\downarrow c} \text{Prob}(d=1 \mid w) - \lim_{w\uparrow c} \text{Prob}\ (d=1 \mid w) = \nabla, \tag{12.2.19}$$

$$P = \lim_{w\uparrow c} \text{Prob}\ (d=1 \mid w).$$

Then

$$\begin{aligned} &\lim_{w\downarrow c}\ E(y \mid w) - \ \lim_{w\uparrow c} E(y \mid w) \\ &\quad = \left\{(P+\nabla)Ey^1 - (1-P-\nabla)Ey^0\right\} - \left[PEy^1 + (1-P)Ey^0\right] \\ &\quad = \nabla E[y^1 - y^0] \end{aligned} \tag{12.2.20}$$

Both the SRD and FRD designs provide only estimates of the ATE at $w_i = c$ for the subgroup of population. The designs do not allow the estimation of the overall ATE.

Let $\psi = \{i \mid w_i < c\}$ and $\bar{\psi} = \{i \mid w_i \geq c\}$, then for the SRD, we may estimate the ATE(c) by the kernel method,

$$\widehat{\text{ATE}}(c) = \frac{\sum\limits_{i\in\bar{\psi}} y_i K(\frac{w_i - c}{h})}{\sum\limits_{i\in\bar{\psi}} K(\frac{w_i - c}{h})} - \frac{\sum\limits_{i\in\psi} y_i K(\frac{w_i - c}{h})}{\sum\limits_{i\in\psi} K(\frac{w_i - c}{h})}, \tag{12.2.21}$$

where $K(\cdot)$ is a kernel function satisfying $K(0) \neq 0, K(v) \to 0$ as $v \to \pm\infty$. Or use the Fan and Gijbels (1992) local linear regression approach

$$\min_{\alpha_0,\beta_0} \sum_{i:c-h<x_i<c} (y_i - \alpha_0 - \beta_0(w_i - c))^2, \tag{12.2.22}$$

and

$$\min_{\alpha_1,\beta_1} \sum_{i:c\leq x_i<c+h} (y_i - \alpha_1 - \beta_1(w_i - c))^2. \tag{12.2.23}$$

Since $E(y^1 \mid w = c) = \hat{\alpha}_1 + \hat{\beta}_1(c - c) = \hat{\alpha}_1$ and $E(y^0 \mid w = c) = \hat{\alpha}_0 + \hat{\beta}_0(c - c) = \hat{\alpha}_0$, therefore,

$$\widehat{\text{ATE}}(c) = \hat{\alpha}_1 - \hat{\alpha}_0. \tag{12.2.24}$$

For FRD,

$$\widehat{\text{ATE}}(c) = \frac{\hat{\alpha}_1 - \hat{\alpha}_0}{\hat{\gamma}_1 - \hat{\gamma}_0}, \tag{12.2.25}$$

where $\hat{\gamma}_1 = \lim\limits_{w\downarrow c} P(d = 1|w)$ and $\hat{\gamma}_0 = \lim\limits_{w\uparrow c} P(d = 1|w)$. They can be obtained as the solution of

$$\min \sum_{i:c\leq x_i<c+h} (d_i - \gamma_1 - \delta_1(w_i - c))^2 \tag{12.2.26}$$

and the solution of

$$\min \sum_{i:c-h\leq x_i<c} (d_i - \gamma_0 - \delta_0(w_i - c))^2. \tag{12.2.27}$$

respectively (for a survey of RD, see Imbens and Lemieux 2008).

12.2.4 Difference-in-Difference (DID) Approach: Imbens and Angrist (1994)

Suppose repeated cross-sectional data are observable at two time periods, t and $t+j$. Suppose for observations in group $\psi = \{i|y_{it} = y^0_{it},\ y_{i,t+j} = y^1_{i,t+j}\}$ and group $\bar{\psi} = \{i|y_{it} = y^0_{it},\ y_{i,t+j} = y^0_{i,t+j}\}$, then DID estimates of ATE are

$$\begin{aligned}\hat{\Delta} &= \left\{E\left(y_{i,t+j}|i \in \psi\right) - E\left(y_{it}|i \in \psi\right)\right\} - \left\{E\left(y_{i,t+j}|i \in \bar{\psi}\right) - E\left(y_{it}|i \in \bar{\psi}\right)\right\} \\ &\simeq \left\{\frac{1}{n_{t+j}}\sum_{i\in\psi} y_{i,t+j} - \frac{1}{n_t}\sum_{i\in\psi} y_{it}\right\} - \left\{\frac{1}{n^*_{t+j}}\sum_{i\in\bar{\psi}} y_{i,t+j} - \frac{1}{n^*_t}\sum_{i\in\bar{\psi}} y_{it}\right\}\end{aligned} \tag{12.2.28}$$

where $n_t = \sum\limits_i 1(y_{it} \in \psi)$, $n_{t+j} = \sum\limits_i 1(y_{i,t+j} \in \psi)$, $n^*_t = \sum\limits_i 1(y_{it} \in \bar{\psi})$, $n^*_{t+j} = \sum\limits_i 1(y_{i,t+j} \in \bar{\psi})$, and 1(A) is an indicator function with 1(A) = 1 and zero otherwise. The first difference provides the average change in the outcomes between t and $t+j$ for $i \in \psi$. However, the first difference of (12.2.28) contains both the treatment effects $E(y^{1*}_{t+j} - y^{0*}_{t+j}|i \in \psi)$ and changes in the external conditions at two different periods, $E(y^{0*}_{t+j} - y^{0*}_t|i \in \psi)$. The second difference of (12.2.28) is to remove the changes due to changes in external conditions through the difference in the changes in control groups $i \in \bar{\psi}$. For instance, the Northern Territory in Australia considered marijuana smoking a criminal act in 1995 but decriminalized it in 1996.[3] The Australian National Drug Strategy Household Surveys provide information about marijuana smoking behavior for residents of the New Territories (NT), New South Wales, Queensland, Victoria, and Tasmania in 1995 and 2001; all except NT were nondecriminalized over this period. The percentage of smokers in NT in 1995 was 0.2342, and in 2001 it was 0.2845. The percentages of residents in nondecriminalized states were 0.1423 in 1995 and 0.1619 in 2001. The difference-in-difference estimate of the impact of decriminalization on marijuana usage is to raise the probability of smoking by (e.g., Damrongplasit et al. 2010)

$$\{(0.2845 - 0.2342) - (0.1619 - 0.1423)\} = 0.0503 \quad -0.0196 = 0.0307.$$

The DID method can provide a valid estimate of the ATE depending on two conditions: (i) no sample selection effect; (ii) $E(y^{0*}_{i,t+j} - y^{0*}_{i,t}|i \in \bar{\psi}) = E(y^{0*}_{i,t+j} - y^{0*}_{i,t}|i \in \psi)$.

We note that $\Delta_{t+j} = E(y^{1*}_{t+j}) - E(y^{0*}_{t+j})$ and $\Delta_t = E(y^{1*}_t) - E(y^{0*}_t)$. The first difference of (12.2.28) is supposed to be an approximation of

$$E(y^{1*}_{t+j}|i \in \psi) - E(y^{0*}_t|i \in \psi) \simeq E(y^{1*}_{t+j} - y^{0*}_{t+j}) + E(y^{0*}_{t+j} - y^{0*}_t). \tag{12.2.29}$$

The second difference of (12.2.28) is supposed to be an approximation of

$$E(y^{0*}_{t+j}) - E(y^{0*}_t). \tag{12.2.30}$$

Hence, for (12.2.28) to get a good estimate of Δ_{t+j}, we need to assume first that there is no sample selection effect, i.e., $\frac{1}{n_{t+j}}\sum\limits_{i\in\psi} y_{i,t+j}$ is a good estimate of $E(y^{1*}_{t+j})$, $\frac{1}{n^*_{t+j}}\sum\limits_{i\in\bar{\psi}} y_{i,t+j}$ is a good estimate of $E(y^{0*}_{t+j})$, and $\frac{1}{n_t}\sum\limits_{i\in\psi} y_{it}$ and that $\frac{1}{n^*_t}\sum\limits_{i\in\bar{\psi}} y_{it}$ are good estimates of $E(y^{0*}_t)$;

[3] Decriminalization does not mean that smoking or possession of small amounts of marijuana is legal. It is still an offense to use or grow marijuana. An individual caught must pay a fine within a specified period to be eligible for the reduced penalty involving no criminal record or imprisonment (e.g., Damrongplasit and Hsiao 2009).

second, $E(y^{0*}_{i,t+j}|i \in \psi) - E(y^{0*}_{it}|i \in \psi)$ can be approximated well by $E(y^{0*}_{i,t+j}|i \in \bar{\psi}) - E(y^{0*}_{it}|i \in \bar{\psi})$.

Under the parametric assumption, the DID method is just the usual dummy-variable approach

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \gamma d_{it} + u_{it}, \tag{12.2.31}$$

where the treatment status dummy (d_{it}) is assumed independent of u_{it}. However, the parametric approach can allow

$$\frac{1}{n_{t+j}} \sum_{i \in \psi} \boldsymbol{x}_{i,t+j} - \frac{1}{n_t} \sum_{i \in \psi} \boldsymbol{x}_{it} \tag{12.2.32}$$

to be different from

$$\frac{1}{n^*_{t+j}} \sum_{i \in \bar{\psi}} \boldsymbol{x}_{i,t+j} - \frac{1}{n^*_t} \sum_{i \in \bar{\psi}} \boldsymbol{x}_{it}. \tag{12.2.33}$$

Further, as long as there is no confounding effects for the treatment status dummy, one can have a more general specification than (12.2.31),

$$y_{it} = \boldsymbol{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{\delta} \cdot 1(d_{it} = 1)) + \gamma d_{it} + u_{it}, \tag{12.2.34}$$

that simultaneously allow changes in intercept and slope coefficients. Then $\gamma = 0$ and/or $\boldsymbol{\delta} = \boldsymbol{0}$ becomes a testable hypothesis.

Summary of Cross-Sectional Approaches

The advantages of the parametric approach are that it can simultaneously take account of both selection on observables and selection on unobservables. It can also estimate the impact of each explanatory variable. The disadvantage is that it needs to impose both functional form and distributional assumptions. If the prior information is inaccurate, the resulting inferences are misleading.

The advantage of the semiparametric approach is that there is no need to prespecify $f(\varepsilon^1, \varepsilon^0, v)$ to take account of the issues of selection on observables and/or unobservarbles. The disadvantage is that it still relies on the correct specifications of $E(y^{1*}|\boldsymbol{x}) = g_1(\boldsymbol{x})$ and $E(y^{0*}|\boldsymbol{x}) = g_0(\boldsymbol{x})$, but some of the critical coefficients may not be estimable (see Chapter 7).

The advantages of the nonparametric approach are that there is no need to impose any assumption on the conditional mean functions of the effects of unobservables. The disadvantages are that some sort of the conditional independence assumption have to hold conditional on some confounding variables. Hence, it only takes account of the issues of selection on observables, and it is not feasible to estimate the impact of each observable factor. In other words, the advantages of the parametric and semiparametric approaches are the disadvantages of the nonparametric approach, and the disadvantages of the parametric and semiparametric approaches are the advantages of the nonparametric approach.

12.3 PANEL DATA APPROACH

Panel data provide information for a number of individuals over time. Over time, some individuals change the treatment status, some not. The information on inter-individual dependence and intra-individual dynamics provides the possibility of relaxing the restrictive conditions needed for the analysis of cross-sectional data as well as to allow the treatment effects to be evolutionary rather than in the form of "either or."

For ease of exposition, we shall assume there are N cross-sectional units, each observed over T periods. We assume that all cross-sectional units did not receive the treatment from $t = 1, \dots, T$, but that the first unit received the treatment from $t = T_1 + 1, \dots, T$, while the other $(N - 1)$ units did not, so $y_{1t} = y_{1t}^{0*}$ for $t = 1, \dots, T_1$ and $y_{1t} = y_{1t}^{1*}$ for $t = T_1 + 1, \dots, T$, while $y_{it} = y_{it}^{0*}$ for $i = 2, \dots, N$ and $t = 1, \dots, T$. That is, the treatment status dummy takes the value $d_{1t} = 0$ for $t = 1, \dots, T_1$, and $d_{1t} = 1$ for $t = T_1 + 1, \dots, T$, while $d_{it} = 0$ for $i = 2, \dots, N$, and $t = 1, \dots, T$.

12.3.1 Nonparametric Method

The treatment effect for the first unit at t is,

$$\Delta_{1t} = y_{1t}^{1*} - y_{1t}^{0*}, \quad t = T_1 + 1, \dots, T. \tag{12.3.1}$$

Because panel data contain information over time, it is possible to also examine the evolution of Δ_{1t} over time. Since $y_{1t} = y_{1t}^{1*}$ for $t = T_1 + 1, \dots, T$, the estimated treat effects $\hat{\Delta}_{1t}$ can be obtained if the predicted value of y_{1t}^{0*} can be obtained. An advantage of using panel data is it allows an investigator to exploit the interrelationships among cross-sectional units to obtain more accurate counterfactuals.

12.3.1.1 Regression Method (PDA)

Let $\boldsymbol{w}_t$ denote all observed variables that are independent of d_{1t} at time t. For simplicity, we let $\boldsymbol{w}_t' = (y_{2t}, \dots, y_{Nt}, \boldsymbol{x}_{1t}', \dots, \boldsymbol{x}_{Nt}')$. Under the assumption that

$$\boldsymbol{w}_t \perp d_{1t}, \tag{12.3.2}$$

we can consider[4]

$$y_{1t}^{0} = y_{1t}^{0*} = E(y_{1t}^{0} | \boldsymbol{w}_t) + \eta_{1t}, \quad t = 1, \dots, T, \tag{12.3.3}$$

where $E(\eta_{1t} | \boldsymbol{w}_t) = 0$. Then $E(y_{1t}^{0} | \boldsymbol{w}_t)$ is an unbiased predictor for y_{1t}^{0*}. Approximating $E(y_{1t}^{0*} | \boldsymbol{w}_t)$ by a linear function of $\boldsymbol{w}_t$, we can write (12.3.3) as[4]

$$y_{1t}^{0*} = a + \boldsymbol{c}' \boldsymbol{w}_t + \eta_{1t}, \quad t = 1, \dots, T. \tag{12.3.4}$$

When T_1 is large, $(a, \boldsymbol{c}')$ can be consistently estimated by the least squares or weighted least squares method. However, in finite T_1, there is a trade-off between the within-sample goodness of fit and the post-sample prediction accuracy. Hsiao, Ching, and Wan (2012) (HCW) propose to fit a subset of $\boldsymbol{w}_t$, say $\boldsymbol{w}_t^*$, for the regression model (12.3.4) using some model selection criterion (e.g., AIC (Akaike 1973), AICC (Hurvich and Tsai 1989), or BIC (Schwarz 1978)). Li and Bell (2017) propose to use LASSO (Tibshirani 1996) to select $\boldsymbol{w}_t^*$. Once $\hat{a}$ and $\hat{\boldsymbol{c}}$ are obtained, HCW propose to construct

$$\hat{y}_{1t}^{0*} = \hat{a} + \hat{\boldsymbol{c}}' \boldsymbol{w}_t^*, \quad t = T_1 + 1, \dots, T. \tag{12.3.5}$$

[4] Hsiao et al. (2012) assume $\boldsymbol{y}_t$ is generated by a factor model and use only y_{jt} in (12.3.4). In practice, any x_{ikt} can be used as an additional predictor as long as they are not affected by the treatment dummy d_{1t}.

and the treatment effects is estimated as

$$\hat{\Delta}_{1t} = y_{1t} - \hat{y}^{0*}_{1t}, \quad t = T_1 + 1, \ldots, T. \tag{12.3.6}$$

Since $y_{1t} = y^{1*}_{1t}$ is observed, the asymptotic variance of $\hat{\Delta}_{1t}$ is just the prediction error variance of y^{0*}_{1t}. If η_{1t} is independently identically distributed over time, the prediction error variance is just

$$\sigma^2_{y^0_{1t}} = \sigma^2_{\eta_1}[1 + (1,\ \boldsymbol{w}^{*'}_t)(W^{*'}W^*)^{-1}(1,\ \boldsymbol{w}^{*'}_t)'], \qquad t = T_1 + 1, \ldots, T. \tag{12.3.7}$$

where $W^* = (\boldsymbol{w}^*_1, \ldots, \boldsymbol{w}^*_{T_1})'$. Hence, the confidence band for Δ_{1t} can be easily constructed as

$$\hat{\Delta}_{1t} \pm c\sigma_{y^0_{1t}}, \tag{12.3.8}$$

where c is chosen by the desired confidence level.

Cross-sectional data provide measurement of policy intervention as a once-and-for-all impact. Panel data allow the policy impact to be evolutionary. If Δ_{1t} is serially correlated but stationary, one can further model the time-varying treatment effects by an autoregressive moving average model using the Box–Jenkins (1970) methodology

$$a(L)\Delta_{1t} = \mu + \theta(L)\eta_t \tag{12.3.9}$$

where L is the lag operator, η_t is an i.i.d. process with zero mean and constant variance, and the roots of $\theta(L) = 0$ lie outside the unit circle. If the roots of $a(L) = 0$ all lie outside the unit circle, the treatment effect is stationary, and the long-term treatment effect is

$$\Delta_1 = a(L)^{-1}\mu = \mu^*. \tag{12.3.10}$$

Alternatively, one can estimate the long-run impact by taking the simple average of $\hat{\Delta}_{1t}$. When both T_1 and $(T - T_1)$ go to infinity,

$$\operatorname*{plim}_{(T-T_1)\to\infty} \frac{1}{T - T_1} \sum_{t=T_1+1}^{T} \hat{\Delta}_{1t} = \Delta_1 \tag{12.3.11}$$

The variance of (12.3.11) can be approximated by the heteroscedastic-autocorrelation consistent (HAC) estimator of Newey and West (1987).

Condition (12.3.2) makes no claim about the relationship between d_{1t} and y^{1*}_{1t} or y^{0*}_{1t}. They can be correlated. All we need is that the y_{jt} are independent of d_{1t} for $j \neq 1$. The approach can be viewed as a "measurement without theory" approach or a nonparametric approach. If $(T - T_1)$ is large, one can also use the above methodology to obtain $\hat{y}^{1*}_{1t} = \hat{\bar{a}} + \hat{\boldsymbol{d}}\tilde{\boldsymbol{w}}_t$, then construct $\hat{\Delta}_t = (\hat{y}^*_{1t} - y^0_{1t})$ for the pretreatment period, $t = 1, \ldots, T_1$ (Fujiki and Hsiao 2015).

12.3.1.2 Synthetic Control Method

Abadie et al. (2010) proposed an alternative panel data approach, which they called the synthetic control method (SCM). The SCM predicts y^{0*}_{1t} by

$$\hat{y}^{0*}_{1t} = \sum_{i=2}^{N} b_i y_{it}, \quad t = T_1 + 1, \ldots, T, \tag{12.3.12}$$

where b_i is obtained by minimizing

$$\left[\begin{pmatrix} \boldsymbol{y}_1 \\ \bar{\boldsymbol{x}}_1 \end{pmatrix} - \begin{pmatrix} Y \\ \bar{X} \end{pmatrix} \boldsymbol{b}\right]' V \left[\begin{pmatrix} \boldsymbol{y}_1 \\ \bar{\boldsymbol{x}}_1 \end{pmatrix} - \begin{pmatrix} Y \\ \bar{X} \end{pmatrix} \boldsymbol{b}\right], \tag{12.3.13}$$

subject to the constraint that

$$b_i \geq 0, \quad \sum_{i=2}^{N} b_i = 1. \tag{12.3.14}$$

where $\boldsymbol{y}_1$ and Y denote the $T_1 \times 1$ and $T_1 \times (N-1)$ matrix of pre-treatment y_{1t} and $y_{jt}, j = 2, \ldots, N$, respectively, $\bar{\boldsymbol{x}_1}$ and $\bar{X}$ denote the pre-treatment time series average of $\boldsymbol{x}_{1t}$ and $\boldsymbol{x}_{jt}, j = 2, \ldots, N$, respectively, and V is a positive definite matrix.

Conditional on y_{jt} independent of d_{1t}, the difference between the PDA and the SCM is that the former is an unconstraint regression while the latter restricts the regression model (12.3.4) intercept $a = 0$ and $\boldsymbol{c}$ to satisfy (12.3.14). If the restrictions are correct, then the SCM is more efficient. If the restrictions are not correct, SCM could lead to a biased prediction of the counterfactuals while the PDA remains unbiased. Gardeazabal and Vega-Bayo (2016) and Wan, Xie, and Hsiao (2018) have provided Monte Carlo studies on the pros and cons of the two approaches.

12.3.2 Parametric Approach

Suppose there is a prior information to specify the data generating process for y_{it}^0 as

$$y_{it}^0 = \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it}. \tag{12.3.17}$$

If $E(v_{it}|\boldsymbol{x}_{it}) = 0$ and $\boldsymbol{x}_{it} \perp d_{it}$, then conditional on $\boldsymbol{\beta}$ and $\boldsymbol{x}_{it}$, an unbiased predictor of y_{it}^0 is

$$\hat{y}_{it}^0 = \boldsymbol{x}_{it}'\hat{\boldsymbol{\beta}}. \tag{12.3.18}$$

However, variables affecting the outcomes are numerous. If $E(v_{it}|\boldsymbol{x}_{it}) \neq 0$, (12.3.18) is a biased predictor. If v_{it} are uncorrelated across i, the popular panel data approach is to decompose $v_{it} = \alpha_i + u_{it}$ and assume $E(u_{it}|\boldsymbol{x}_{it}) = 0$. Then the coefficients $\boldsymbol{\beta}$ can be estimated by the covariance method discussed in Chapter 2 if there is no selection bias. If $E(v_{it}|d_{it}) \neq E(v_{it})$, and the individual-specific effects take an additive form, then a sample selection model of the form (7.3.1) and (7.3.4) can be constructed and the parametric or semiparametric methods discussed in Section 7.3 can be used to estimate $\boldsymbol{\beta}$.

If there is no selection on unobservarbles (or no sample selection effects) but v_{it} are cross- correlated and correlated with $\boldsymbol{x}_{it}$, Xu (2017) suggests approximating v_{it} by

$$v_{it} = \boldsymbol{b}_i'\boldsymbol{f}_t + u_{it} \tag{12.3.19}$$

where $\boldsymbol{b}_i'$ are $1 \times r$ individual specific but time-invariant constants, and $\boldsymbol{f}_t'$ are $1 \times r$ individual invariant but time-varying constants. The $\boldsymbol{b}_i$ and/or $\boldsymbol{f}_t$ are allowed to be correlated with $\boldsymbol{x}_{it}$, but $E(u_{it}|\boldsymbol{x}_{it}) = 0$. Under the assumption that y_{it}^0 are generated by (12.3.17) and (12.3.19), conditional on $\boldsymbol{x}_{it}, \boldsymbol{\beta}, \boldsymbol{b}_i$ and $\boldsymbol{f}_t$, the unbiased predictor of y_{it}^0 is

$$\hat{y}_{it}^0 = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{b}_i'\boldsymbol{f}_t. \tag{12.3.20}$$

Given $(y_{it}, \boldsymbol{x}_{it})$, $i = 1, \ldots, N$, and $t = 1, \ldots, T_1$, conditional on r, the unknown $\boldsymbol{\beta}, \boldsymbol{b}_i$ and $\boldsymbol{f}_t$ can be estimated using Bai's (2009a) least squares method (see Chapter 10) based on data of the pre-treatment period, $t = 1, \ldots, T_1$. To obtain post-treatment estimates of $\boldsymbol{f}_t$

for $t = T_1 + 1, \ldots, T$, under the assumption that $\boldsymbol{x}_{it}$ are orthogonal to the treatment status dummy d_{1t}, Xu (2017) suggests a least squares estimator using the post-treatment data of $i = 2, \ldots, N$, and $t = T_1 + 1, \ldots, T$ conditional on the estimated $\boldsymbol{\beta}$ and $\boldsymbol{b}_i, \hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}_i$,

$$\min_{\boldsymbol{f}_t} \sum_{i=2}^{N} (y_{it} - \boldsymbol{x}'_{it}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{b}}'_i \boldsymbol{f}_t)^2, \tag{12.3.21}$$

yields

$$\hat{\boldsymbol{f}}_t = \left(\sum_{i=2}^{N} \hat{\boldsymbol{b}}_i \hat{\boldsymbol{b}}'_i \right)^{-1} \left(\sum_{i=2}^{N} \hat{\boldsymbol{b}}_i (y_{it} - \boldsymbol{x}'_{it}\hat{\boldsymbol{\beta}}) \right), \quad t = T_1 + 1, \ldots, T. \tag{12.3.22}$$

However, the dimension of common factors $\boldsymbol{f}_t$ is usually unknown. Xu (2017) suggests a cross-validation method with steps of the algorithm as:

(a) Starting with the given $\hat{r}$, estimate $\hat{\boldsymbol{\beta}}$ and $\hat{F} = (\hat{\boldsymbol{f}}_{\mathbf{1}}, \ldots, \hat{\boldsymbol{f}}_{T_1})$ from the control group data that goes through all T_1 periods:

(b) In round $s \in \{1, \ldots, T_1\}$, hold back data for all treated units at time s. Run an OLS using the rest of the pretreated data, obtaining factor loads for each treated unit i:

$$\hat{\boldsymbol{b}}_{i,-s}(r) = \left(\hat{F}'_{-s}(r) \hat{F}_{-s}(r) \right)^{-1} \hat{F}'_{-s}(r) \left(Y_{i,-s} - X'_{i,-s}\hat{\boldsymbol{\beta}}(r) \right), \quad i = 1, \ldots, N,$$

in which the subscript “–s” stands for all pretreatment periods except for s.

(c) Predict the treated outcomes at time s using

$$\hat{y}^{(0)}_{is}(r) = \boldsymbol{x}'_{is}\hat{\boldsymbol{\beta}}(r) + \hat{\boldsymbol{b}}'_{i,-s}(r)\hat{\boldsymbol{f}}_s(r), \tag{12.3.23}$$

(d) Obtain the mean square prediction error

$$MSPE(r) = \sum_{t=1}^{T_1} \sum_{i=1}^{N} e^2_{it}(r)/T_1, \tag{12.3.24}$$

where $e^2_{it}(r) = y^0_{it} - \hat{y}^0_{it}(r)$.

Repeat steps (a)–(d) for each r, choosing the $\hat{r}$ that has smallest $MSPE(r)$. Then estimate $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{b}}_i$, and $\hat{\boldsymbol{f}}_t$ as described above. Substituting the identified $\hat{r}$, $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{b}}_1$, and $\hat{\boldsymbol{f}}_{\boldsymbol{t}}$, one can predict y^{0*}_{it} by

$$\hat{y}^{0*}_{1t} = \boldsymbol{x}'_{1t}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{b}}'_1 \hat{\boldsymbol{f}}_t, \quad t = T_1 + 1, \ldots, T. \tag{12.3.25}$$

Substituting the parametrically estimated $\hat{y}^{0*}_{1t}$ in lieu of y^{0*}_{1t} for $t = T_1 + 1, \ldots, T$, one obtains the estimated treatment effect $\hat{\Delta}_{1t} = y_{1t} - \hat{y}^{0*}_{1t}$.

12.3.3 Semiparametric Approach

Conditional on $\boldsymbol{x}_{it}$ and $\boldsymbol{\beta}$, one can obtain

$$v_{it} = y_{it} - \boldsymbol{x}'_{it}\boldsymbol{\beta}. \tag{12.3.26}$$

The semiparametric approach is to obtain $\hat{v}_{1t} = E(v_{1t}|v_{2t}, \ldots, v_{Nt})$ using the nonparametric method. For instance, Hsiao and Zhou (2019) suggest first using the Bai (2009) least squares method to obtain $\hat{\boldsymbol{\beta}}$, then to obtain $\hat{v}_{it} = y_{it} - \boldsymbol{x}_{it}\hat{\boldsymbol{\beta}}$. From $(\hat{v}_{1t}, \ldots, \hat{v}_{Nt})$ apply the nonparametric method described in section 12.3.1 to obtain

$$\hat{v}_{1t} = \hat{\boldsymbol{a}}'\hat{\boldsymbol{v}}^*_t + \hat{\eta}_{1t}, \quad t = 1, \ldots, T. \tag{12.3.27}$$

where $\hat{\boldsymbol{v}}_t^*$ denotes a subset of $(\hat{v}_{2t}, \ldots, \hat{v}_{Nt})$. Based on $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{a}}$, the counterfactuals y_{1t}^{0*} are constructed as

$$\hat{y}_{1t}^{0*} = \boldsymbol{x}_{1t}'\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{a}}'\hat{\boldsymbol{v}}_t^*, \quad t = T_1 + 1, \ldots, T. \tag{12.3.28}$$

12.3.4 Averaging Method

The above sections discussed parametric, semiparametric, and nonparametric ways of constructing counterfactuals. Each has its advantages and disadvantages. Unfortunately, neither the conventional hypothesis testing approach nor the predictive approach (e.g., Diebold and Mariano 1995, White 2001) appears feasible in terms of assessing which method is more likely to generate more accurate counterfactuals in a given situation, because the potential outcomes y_{it}^{1*} or y_{it}^{0*} are unobserved. In practice, no method can claim its superiority over others. The only way to assess which method is more likely to produce more accurate estimates of the treatment effects is through simulation. Hsiao and Zhou (2019) conducted a number of simulations designs to assess the pros and cons of different methods. Based on the criteria of the mean of the absolute bias and the root mean square prediction error at each post-treatment data point, they find that (i) the unknown data generating process matters; (ii) the relative sample size matters; and (iii) no method is able to dominate in a wide array of situations.

Bates and Granger (1969) have argued that even the most complicated model is likely to be misspecified, and that combining forecasting across different models is a way to make the forecast more robust against misspecification biases and measurement errors in the data. Many authors have suggested different methods with which to combine forecasts (e.g., see the survey by Timmerman 2006). Most of these methods depend on the relation between the actuals and forecasts, while in our case the actuals are unobservable. On the other hand, the simulation and an empirical example analyzed by Hsiao and Wan (2014) appear to indicate that no combination method is able to yield more accurate forecasts uniformly over time in a wide array of situations. Both mean-corrected and mean-and-scale-corrected simple average methods appear to be robust ways of combining forecasts. Therefore, Hsiao and Zhou (2019) suggest the following two ways to combine the different methods of generating counterfactuals.

Let $\bar{y}_t = \frac{1}{M}\sum_{j=1}^{M}\hat{y}_{jt}$, where $\hat{y}_{jt}$ denote the within-sample or post-sample predicted value of y_{1t} based on the jth method.

The mean corrected simple average method (MA) is

$$\hat{y}_{1t} = a + \bar{y}_t, \; t = T_1 + 1, \ldots, T, \tag{12.3.29}$$

where a is estimated by

$$a = \frac{1}{T_1}\sum_{t=1}^{T_1}(y_{1t} - \bar{y}_t), \; t = 1, \ldots, T_1. \tag{12.3.30}$$

The mean and scale corrected simple average method (MB) is

$$\hat{y}_{1t} = a + b\bar{y}_t, \; t = T_1 + 1, \ldots, T, \tag{12.3.31}$$

where a and b are obtained by minimizing

$$\sum_{t=1}^{T_1}(y_{1t} - a - b\bar{y}_t)^2, \quad t = 1, \ldots, T_1. \tag{12.3.32}$$

12.4 SOME EXAMPLES

12.4.1 Measuring the Impact of the Closer Economic Partnership Arrangement (CEPA) on Hong Kong

Hong Kong signed the CEPA with Mainland China in June 2003 and started implementing its arrangement in January 2004. The CEPA aimed to strengthen the linkage between Mainland China and Hong Kong by allowing Chinese citizens to enter Hong Kong as individual tourists and liberalizing trade in services, enhancing cooperation in the area of finance, promoting trade and investment facilitation, and mutual recognition of professional qualifications. The implementation of CEPA started on January 1, 2004, when 273 types of Hong Kong products could be exported to the mainland tariff free, another 713 types on January 1, 2005, 261 on January 1, 2006, and a further 37 on January 2007. Chinese citizens residing in selected cities were also allowed to visit Hong Kong as individual tourists, from 4 cities in 2003 to 49 cities in 2007, covering all 21 cities in Guangdong Province.

Hsiao, Ching, and Wan (2012) tried to assess the impact of economic integration of Hong Kong with Mainland China on Hong Kong's economy by comparing what actually happened to Hong Kong's real GDP growth rates with what would have been if there were no CEPA with Mainland China in 2003. More specifically, they analyzed how these events have changed Hong Kong's growth rate.

Because Hong Kong, by comparison, is a tiny city relative to other countries and regions, Hsiao, Ching, and Wan (2012) believe whatever happened in Hong Kong will have no bearing on other countries. In other words, they expect (12.3.2) to hold. Therefore, they use the quarterly real growth rate of Australia, Austria, Canada, China, Denmark, Finland, France, Germany, Indonesia, Italy, Japan, Korea, Malaysia, Mexico, Netherlands, New Zealand, Norway, Philippines, Singapore, Switzerland, Taiwan Thailand, the UK, and the U.S. to predict the quarterly real growth rate of Hong Kong in the absence of intervention. All the nominal GDP, CPI are from the Organization for Economic Cooperation and Development (OECD) Statistics, International Financial Statistics and CEIC Database.

Using the AICC criterion, the countries selected were Austria, Italy, Korea, Mexico, Norway, and Singapore. OLS estimates of the weights are reported in Table 12.1. Actual and predicted growth paths from 1993Q1 to 2003Q4 are plotted in Figure 12.1. The

Table 12.1. *AICC selected model using data for the period 1993Q1–2003Q4*

	Beta	Std	T
Constant	−0.0019	0.0037	−0.524
Austria	−1.0116	0.1682	−6.0128
Italy	−0.3177	0.1591	−1.9971
Korea	0.3447	0.0469	7.3506
Mexico	0.3129	0.051	6.1335
Norway	0.3222	0.0538	5.9912
Singapore	0.1845	0.0546	3.3812
R2 = 0.931			
AICC = −378.9427			

Source: Hsiao et al. (2012, Table 20).

Table 12.2. *Treatment effect for economic integration 2004Q1–2008Q1 based on AICC selected model*

	Actual	Control	Treatment
Q1-2004	0.077	0.0493	0.0277
Q2-2004	0.12	0.0686	0.0514
Q3-2004	0.066	0.0515	0.0145
Q4-2004	0.079	0.0446	0.0344
Q1-2005	0.062	0.0217	0.0403
Q2-2005	0.071	0.0177	0.0533
Q3-2005	0.081	0.0333	0.0477
Q4-2005	0.069	0.029	0.04
Q1-2006	0.09	0.0471	0.0429
Q2-2006	0.062	0.0417	0.0203
Q3-2006	0.064	0.025	0.039
Q4-2006	0.066	0.0009	0.0651
Q1-2007	0.055	−0.0101	0.0651
Q2-2007	0.062	0.0092	0.0528
Q3-2007	0.068	0.0143	0.0537
Q4-2007	0.069	0.0508	0.0182
Q1-2008	0.073	0.0538	0.0192
MEAN	0.0726	0.0323	0.0403
STD	0.0149	0.0213	0.016
T	4.8814	1.5132	2.5134

Source: Hsiao et al. (2012, Table 21).

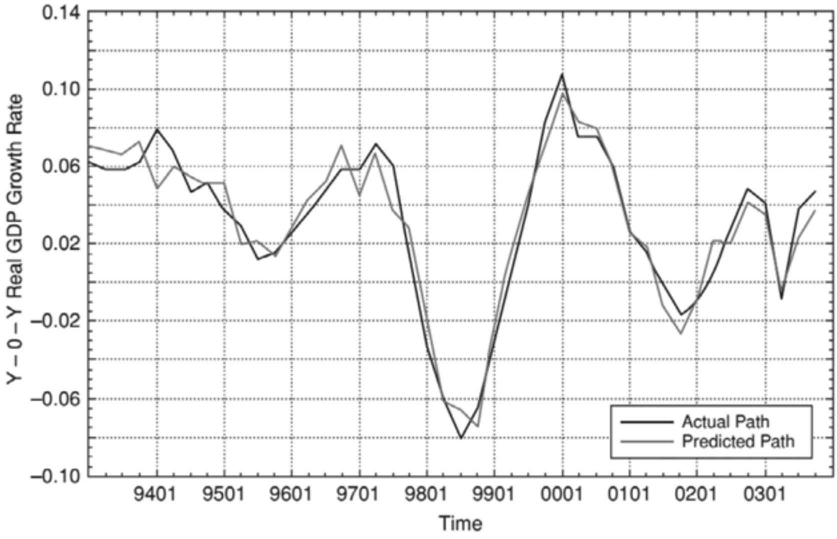


Figure 12.1. Actual and AICC predicted real GDP growth rate from 1993Q1 to 2003Q4. *Source:* Hsiao et al. (2012, Fig. 7)

estimated quarterly treatment effects are reported in Table 12.2. The actual and predicted counterfactual for the period 2004Q1 to 2008Q1 are presented in Figure 12.2. Using the AIC criterion, the selected group consists of Austria, Germany, Italy, Korea, Mexico, Norway, Philippines, Singapore, and Switzerland. The OLS estimates of the weights are

Table 12.3. *AIC selected model using data for the period 1993Q1–2003Q4*

	Beta	Std	T
Constant	−0.0030	0.0042	−0.7095
Austria	−1.2949	0.2181	−5.9361
Germany	0.3552	0.233	1.5243
Italy	−0.5768	0.1781	−3.2394
Korea	0.3016	0.0587	5.1342
Mexico	0.2340	0.0609	3.8395
Norway	0.2881	0.0562	5.1304
Switzerland	0.2436	0.1729	1.4092
Singapore	0.2222	0.0553	4.0155
Philippines	0.1757	0.1089	1.6127
R2 = 0.9433			
AIC = −385.7498			

Source: Hsiao et al. (2012, Table 22).

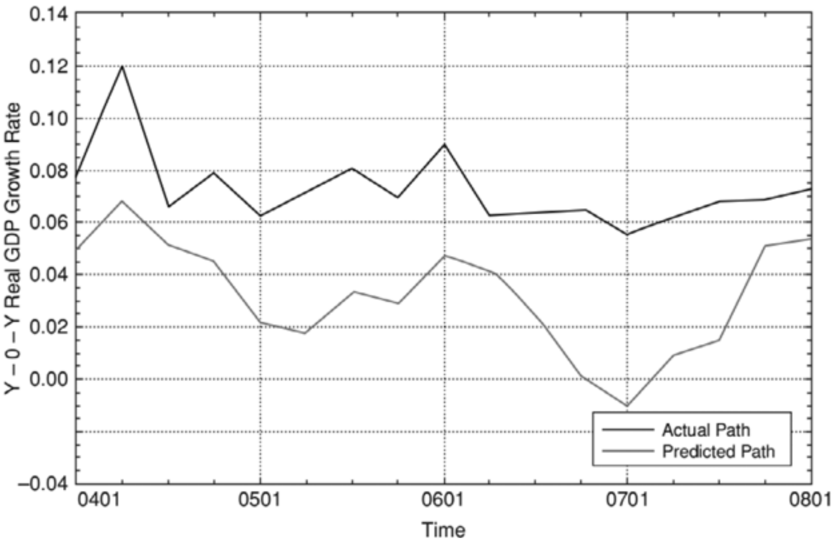


Figure 12.2. AICC – Actual and counterfactual real GDP growth rate from 2004Q1 to 2008Q1.
Source: Hsiao et al. (2012, Fig. 8)

in Table 12.3, and the estimated quarterly treatment effects are in Table 12.4. The pre- and post-intervention actual and predicted outcomes are plotted in Figures 12.3 and 12.4.

It is notable that even though the two models use different combinations of countries, both groups of countries trace closely the actual Hong Kong path before the implementation of CEPA (with R^2 above 0.93). It is also quite remarkable that the post-sample predictions closely matched the actual turning points at a lower level for the treatment period even though no Hong Kong data were used. The CEPA effect at each quarter was all positive and appeared to be serially uncorrelated. The average actual growth rate from 2004Q1 to 2008Q1 is 7.26%. The average projected growth rate without CEPA is 3.23% using the group of countries selected by AICC and 3.47% using the group selected by

Table 12.4. *AIC – Treatment effect for economic integration 2004Q1–2008Q1 based on AIC selected model*

	Actual	Control	Treatment
Q1-2004	0.077	0.0559	0.0211
Q2-2004	0.12	0.0722	0.0478
Q3-2004	0.066	0.0446	0.0214
Q4-2004	0.079	0.0314	0.0476
Q1-2005	0.062	0.0121	0.0499
Q2-2005	0.071	0.0126	0.0584
Q3-2005	0.081	0.0314	0.0496
Q4-2005	0.069	0.0278	0.0412
Q1-2006	0.09	0.0436	0.0464
Q2-2006	0.062	0.0372	0.0248
Q3-2006	0.064	0.0292	0.0348
Q4-2006	0.066	0.0122	0.0538
Q1-2007	0.055	0.0051	0.0499
Q2-2007	0.062	0.0279	0.0341
Q3-2007	0.068	0.0255	0.0425
Q4-2007	0.069	0.0589	0.0101
Q1-2008	0.073	0.062	0.011
Mean	0.0726	0.0347	0.0379
Std	0.0149	0.0193	0.0151
T	4.8814	1.7929	2.5122

Source: Hsiao et al. (2012, Table 23).

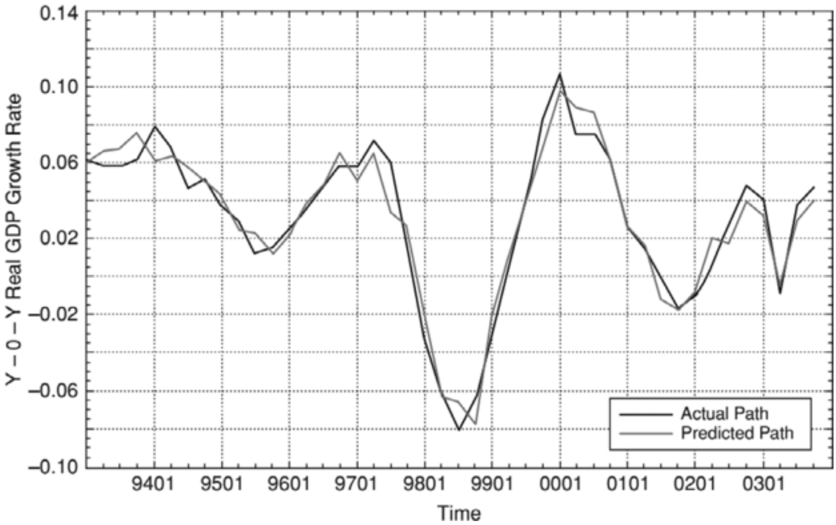


Figure 12.3. Actual and AIC predicted real GDP growth rate from 1993Q1 to 2003Q4. *Source:* Hsiao et al. (2012, Fig. 10)

AIC. The estimated average treatment effect is 4.03% with a standard error of 0.016 based on the AICC group, and 3.79% with a standard error of 0.0151 based on the AIC group. The t-statistic is 3.5134 for the former group and 3.5122 for the latter group. Either set of countries yields similar predictions and highly significant CEPA effects. In other words,

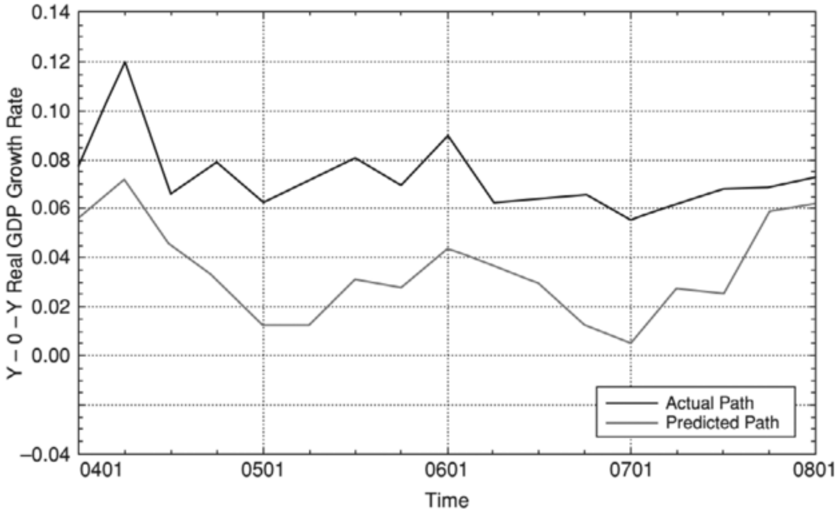


Figure 12.4. AIC – Actual and counterfactual real GDP growth rate from 2004Q1 to 2008Q1.
Source: Hsiao et al. (2012, Fig. 11)

through liberalization and increased openness with Mainland China, the real GDP growth rate of Hong Kong was raised by more than 4% compared to the growth rate had there been no CEPA agreement with Mainland China.

12.4.2 Measuring the Impact of California's Tobacco Control Program (CTCP)

In November 1988, California passed Proposition 99, which increased California's cigarette tax by 25 cents per pack and earmarked the tax revenue for health and antismoking measures. Proposition 99 triggered a wave of local clean-air ordinances in California. Abadie et al. (2010) used the synthetic control method (SCM) for the period 1970–2000 to show that not only did the California Tobacco Control Program (CTCP) have a significant impact on per capita cigarette consumption for the period 1989–2000, but its impact continued to be enhanced over time. Hsiao and Zhou (2019) revisited the effectiveness of CTCP on per capita cigarette consumption and personal healthcare expenditures using the panel parametric, semiparametric, nonparametric, and model averaging methods discussed in Section 12.3.

In addition to state per capita cigarette consumption, Abadie et al. (2010) also used per capita GDP, the price of cigarettes, per capita beer consumption, and the percentage of the population aged 15–24 as observable factors that might affect cigarette consumption. However, apart from per capita GDP and the percentage of the population aged 15–24, the other variables were likely to be affected by the policy treatment, thus violating (12.3.2).[5] On the other hand, the National Institute on Drug Abuse (NIDA) (2018) claims that people living below the poverty line and those with low education attainment are more likely to smoke than those in the general population and that they are not likely to be affected

[5] Hsiao and Zhou (2019) did not include the per-capita beer consumption, as in Abadie et al. (2010), because these data contain many unrecoverable missing observations. The data of personal healthcare expenditures is collected from the National Health Service (NHS) website.

Table 12.5. *Comparison of the actual and counterfactual cigarette consumption*

Year	Actual	SCM	PCA	CCE	CPDA	PDA	PDAX	MA	MB
1989	82.4	88.8	81.3	86.4	86.8	89.7	88.7	86.6	85.9
1990	77.8	86.7	72.9	81.5	81.9	84.9	84.2	81.0	80.2
1991	68.7	81.8	68.1	74.3	75.2	78.4	81.5	75.5	74.6
1992	67.5	81.3	64.9	72.3	72.4	76.6	80.0	73.2	72.3
1993	63.4	81.1	59.2	70.9	70.7	75.3	80.3	71.3	70.3
1994	58.6	80.7	51.7	65.3	67.9	73.2	76.8	67.0	65.9
1995	56.4	78.0	47.6	65.9	67.9	73.0	76.1	66.1	64.9
1996	54.5	77.1	42.9	66.7	65.8	70.6	74.3	64.0	62.9
1997	53.8	77.3	39.8	66.2	65.1	70.3	73.7	63.0	61.8
1998	52.3	73.7	40.6	64.2	64.6	71.7	70.7	62.4	61.2
1999	47.2	73.1	35.5	61.9	62.7	69.2	68.9	59.6	58.4
2000	41.6	66.8	29.6	57.5	57.5	62.9	63.2	54.1	52.7
MAB		18.5	7.46	9.12	9.56	14.3	16.2	8.33	7.27

Notes: MAB denotes the mean absolute impact, and "SCM" refers to the counterfactuals replicated from Abadie et al. (2010).
Source: Hsiao and Zhou (2019).

by cigarette consumption. Thus, Hsiao and Zhou (2019) use the variables poverty rate and education attainment, in lieu of the other variables used by Abadie et al. (2010), as observable exogenous factors and treat all other factors that could affect cigarette consumption as unobservable factors in the counterfactual analysis.

Hsiao and Zhou (2019) followed Abadie et al. (2010) in excluding the four states (Arizona, Florida, Massachusetts, and Oregon) with other large-scale tobacco control programs and seven states (Alaska, Hawaii, Maryland, Michigan, New Jersey, New York, and Washington) that raised their state cigarette taxes by 50 cents or more over the 1989–2000 period. The District of Columbia is also excluded from our sample. Thus, their control group includes 38 states. The data is collected between 1970 and 2000 for cigarette consumption and between 1980 and 2000 for personal healthcare expenditures.

In Hsiao and Zhou's (2019) sample, N $(= 38)$ was greater than T_1 ($T_1 = 19$ for cigarette consumption, $T_1 = 9$ for personal healthcare expenditures). For the nonparametric and semiparametric methods to select a subset of control variables, they use the LASSO method to select the subset of control units with which to generate counterfactuals.

Table 12.5 provides the actual and estimated counterfactuals for cigarette consumption per capita based on the parametric method (PCA), the semiparametric method using a complete set of control variables (CCE), the semiparametric method using LASSO to select a subset of control units (CPDA), the nonparametric method using $\boldsymbol{y}_j$ as control variables (PDA), and the nonparametric method with control units consisting of both $\boldsymbol{y}_j$ and $\boldsymbol{x}_j$ (PDAX) as well as model averaging methods MA and MB. Figure 12.5 plots the difference between the actual outcomes and the counterfactuals.[6] As one can see, in general, the

[6] The PCA approach yields unintuitive positive treatment effects. This could be due to misspecification of the parametric model or/and inability of the factor structure to fully capture the omitted individual-time varying effects. Or it could be due to the conventional overemphasis of California cigarette tax effects. The decline in per-capita cigarette consumption could be due to other factors, say the success of high school education on the harm of cigarette consumption and political pressure, etc. No one really knows the actual treatment effects because counterfactuals, by definition, mean something never happened. Furthermore, because CPDA, PDAX rely on the same set of covariates, Hsiao and Zhou opt to keep the results obtained by PCA as they are, although they appear counterintuitive.

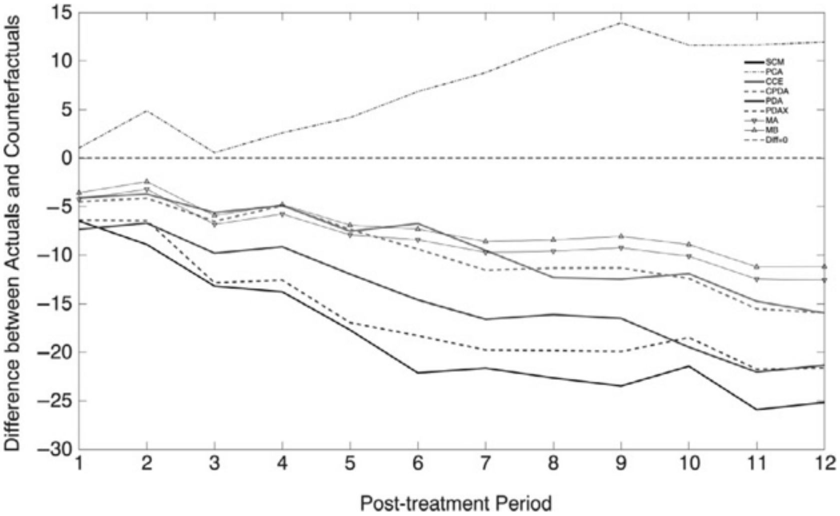


Figure 12.5. Difference between the actual and counterfactual cigarette consumption in the treatment period.
Source: Hsiao and Zhou (2019)

SCM and the PDAX provide the lower bounds of the counterfactuals, the PCA provides the upper bound, and CCE and CPDA (the method that appears to yield counterfactuals closest to the unknown DGP in our simulation) yield counterfactuals that are in the middle. In general, their findings confirmed Abadie et al.'s (2010) finding that Proposition 99 had a significant treatment effect on cigarette consumption. However, the absolute magnitude of the treatment effects due to the 1988 tax increase might be not as large as estimated initially, and often less than half. Neither the estimate based on MA nor that based on MB shows any increasing trend of treatment effects.

Since counterfactuals are unobservable, it is hard to know which estimates are close to the true treatment effects. Therefore, Hsiao and Zhou (2019) proposed to use indirect evidence to gauge which estimates could be closer to the actual treatment effects. Cigarette smoking has been linked to about 80%–90% of all cases of lung cancer. In addition, smoking causes lung diseases such as bronchitis and emphysema and increases the risk of heart disease. Smoking is also linked to many other major health conditions, including diabetes, rheumatoid arthritis, inflammation, and impaired immune function (e.g., Moore 1996, National Institute of Drug Abuse (NIDA) 2018). Given the established medical link between cigarette consumption and health conditions, Hsiao and Zhou (2019) assumed studying California residents' personal healthcare expenditures could shed light on the impact of CTCP on cigarette consumption. They note that, over time, health expenditures could be affected by many socioeconomic factors and advancement in medical science. However, as long as these factors affect all states' health expenditures in more or less the same way, using cross-sectional state health expenditures might capture these unobserved effects.

To avoid the outcomes being contaminated by the treatment, again, as in the study of cigarette consumption, Hsiao and Zhou (2019) use only personal healthcare expenditures for the 38 states from 1980 and 2000 and include per capita income as an additional control variable. Table 12.6 provides the actual and estimated counterfactuals for personal healthcare expenditures during the treatment period. As we can see from Table 12.6, all of the

Table 12.6. *Comparison of the actual and counterfactual personal healthcare expenditures*

Year	Actual	SCM	PCA	CCE	CPDA	PDA	PDAX	MA	MB
1989	8.994	8.927	8.974	8.981	8.976	8.794	8.974	8.976	8.984
1990	9.120	9.029	9.084	9.092	9.082	9.088	9.088	9.087	9.097
1991	9.241	9.116	9.174	9.184	9.174	9.188	9.188	9.182	9.194
1992	9.317	9.187	9.244	9.247	9.229	9.263	9.263	9.249	9.263
1993	9.382	9.243	9.299	9.305	9.281	9.325	9.325	9.307	9.322
1994	9.441	9.297	9.346	9.354	9.325	9.370	9.370	9.353	9.368
1995	9.506	9.362	9.417	9.428	9.386	9.424	9.424	9.416	9.432
1996	9.568	9.403	9.478	9.487	9.436	9.463	9.463	9.465	9.483
1997	9.635	9.460	9.544	9.558	9.504	9.530	9.530	9.533	9.552
1998	9.658	0.523	9.609	9.628	9.564	9.589	9.589	9.596	9.615
1999	9.693	9.578	9.679	9.669	9.623	9.649	9.649	9.660	9.681
2000	9.755	9.637	9.765	9.789	9.702	9.709	9.709	9.735	9.757
MAB		0.128	0.059	0.053	0.085	0.061	0.061	0.062	0.047

Source: Hsiao and Zhou (2019).

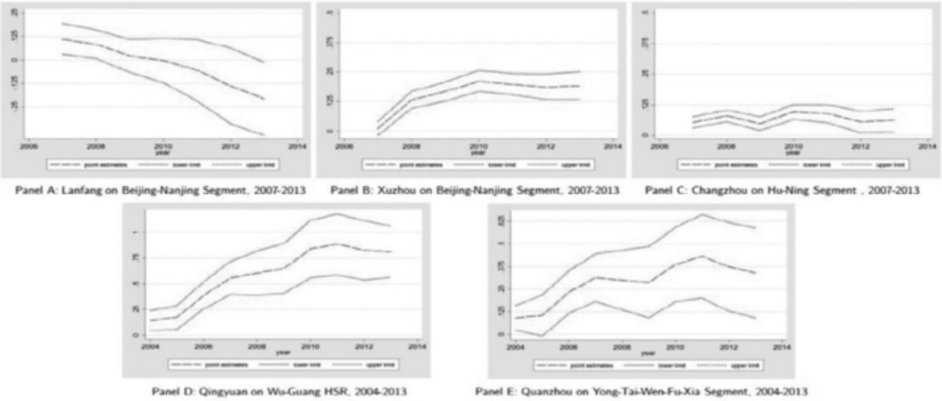
different methods yield similar counterfactuals. Contrary to the estimates of treatment effects on per capita cigarette consumption increasing (in absolute value) over time (Table 12.5), there are hardly any treatment effects on personal health expenditures between actual and counterfactuals after 10 or more years of the implementation of the 1989 CTCP. The 1999 actual is 9.693, and the model average counterfactual estimate is 9.660. Likewise, the 2000 actual is 9.755, and the model average counterfactual estimate is 9.735.

Using the healthcare expenditures information as corroborating evidence, it appears that, although there appears to be a discouraging effect of the CTCP on per capita cigarette consumption, contrary to the common belief, the CTCP's absolute impact may not be increasing over time and could be close to the MA or MB estimates.

12.5 MULTIPLE TREATED UNITS

The panel data approach discussed in Section 12.3 is concerned with measuring the treatment effects on an individual over time. However, a decision maker could probably be more interested in some measures of the average treatment effects. When there are multiple treated units, it provides the possibility of measuring the average treatment effects at a given time, ATE(t) defined in (12.1.2) and/or the average treatment effects defined in (12.1.3).

However, individual treatment effects may be heterogeneous. For instance, Figure 12.6 reproduces the Ke et al. (2017) plots of the changes in per capita GDP growth rate on some Chinese localities over time after a high-speed rail line was constructed. As one can see, the treatment effects on different localities are quite different. This raises the issue of whether the differences in treatment effects are due to some fundamental differences in the localities or just due to the working of some chance mechanism (i.e., the different treatment effects can be viewed as random draws from a common population). If the treatment effects are heterogeneous, does it make sense to discuss ATE? If the treatment effects are homogeneous, what is the best way to estimate ATE or aggregate individual treatment effects?



Panel A: Lanfang on Beijing-Nanjing Segment, 2007-2013

Panel B: Xuzhou on Beijing-Nanjing Segment, 2007-2013

Panel C: Changzhou on Hu-Ning Segment , 2007-2013

Panel D: Qingyuan on Wu-Guang HSR, 2004-2013

Panel E: Quanzhou on Yong-Tai-Wen-Fu-Xia Segment, 2004-2013

Figure 12.6. The impact of China's high speed rail projects on the per capita growth rate of some Chinese cities.
Source: Ke et al. (2017).

12.5.1 "Homogeneous" or "Heterogeneous" Treatment Effects

Treatment effects are considered *homogeneous* if the difference in treatment effects between cross-sectional units can be considered as due to a chance mechanism, i.e.,

$$E(\Delta_{it}|\boldsymbol{x}_{it}) = \Delta_t = E(\Delta_{it}) \tag{12.5.1}$$

and *heterogeneous* if

$$E(\Delta_{it}|\boldsymbol{x}_{it}) \neq \Delta_t, \tag{12.5.2}$$

where $\boldsymbol{x}_{it}$ are some conditional covariates. Then a test of homogeneous treatment effects can be put in a regression framework,

$$\hat{\Delta}_{it} = \lambda_t + \boldsymbol{x}_{it}'\boldsymbol{\delta} + u_{it}, \quad \begin{array}{l} i = 1, \ldots, N. \\ t = 1, \ldots, T. \end{array} \tag{12.5.3}$$

The null of homogeneous versus heterogeneous treatment effects at time t is to test the null

$$H_0 : \boldsymbol{\delta} = 0. \tag{12.5.4}$$

against the alternative

$$H_1 : \boldsymbol{\delta} \neq 0. \tag{12.5.5}$$

An F-test can easily be constructed for (12.5.3). Similarly, a test of homogeneity over i and t is an F-test of

$$H_0' : \boldsymbol{\lambda} = 0,\ \boldsymbol{\lambda}_t = 0 \text{ for } t = T_1 + 1, \ldots, T \tag{12.5.6}$$

versus

$$H_1' : H_0' \text{ is not true.} \tag{12.5.7}$$

For instance, in the study of the impact of China's high-speed rail projects on the local's per capita GDP growth rate, Ke et al. (2017) used sectorial output or employment share, local infrastructure, index of human capital measure, and index of local tourism measure, city size, etc., as the control covariates $\boldsymbol{x}$ and rejected homogeneity.

12.5.2 Aggregate or Disaggregate Measure of Treatment Effects

A test of homogeneous treatment effects is important in understanding the estimated ATE. When the treatment effects are heterogeneous, it raises the issue of how to interpret measured ATE (e.g., Hsiao et al. 2005; Lee and Pesaran 1993; Pesaran et al. 1994; Theil 1954; Van Garderen 2000).

12.5.2.1 Aggregate Measure

We consider an aggregate measure (ATE) as

$$\Delta_t = \sum_{i=1}^{N} w_i \Delta_{it}. \tag{12.5.8}$$

where

$$w_i \geq 0, \quad \sum_{i=1}^{N} w_i = 1. \tag{12.5.9}$$

There are many aggregation methods that satistfy (12.5.8) and (12.5.9). For instance, the conventional simple average is to let $w_i = \frac{1}{N}$. However, one could also consider aggregating individual outcomes using the weight $w_i = w_i^{*2}$ or w_i^{**2} where $\boldsymbol{w}^* = (w_1^*, \ldots, w_N^*)'$ or $\boldsymbol{w}^{**} = (w_1^{**}, \ldots, w_N^{**})'$ is the eigenvector corresponding to the smallest and the largest eigenvalue of the determinant equation,

$$\left| \frac{1}{(T - T_1)} \sum_{t=T_1+1}^{T} (\boldsymbol{\Delta}_t - \bar{\boldsymbol{\Delta}})(\boldsymbol{\Delta}_t - \bar{\boldsymbol{\Delta}})' - \delta\, I_N \right| = 0 \tag{12.5.10}$$

where $\boldsymbol{\Delta}_t = (\Delta_{1t}, \ldots, \Delta_{Nt})'$, $\bar{\boldsymbol{\Delta}} = \frac{1}{(T-T_1)} \sum_{t=T_1+1}^{T} \boldsymbol{\Delta}_t$. All the weights in (12.5.8)–(12.5.10) are nonnegative and sum to 1. Let Δ_t, Δ_t^* and Δ_t^{**} denote (12.5.8) using $\frac{1}{N}, w_i^{*2}$ and w_i^{**2}, respectively. Then $\Delta_t \neq \Delta_t^* \neq \Delta_t^{**}$. They have different interpretations of the ATE(t) and

$$\text{Var}(\Delta_t^*) \leq \text{Var}(\Delta_t) \leq \text{Var}(\Delta_t^{**}). \tag{12.5.11}$$

Considering decomposing y_{it} into the sum of the long-run component, μ_{it} and the transitory component, v_{it}, then method of using w_i^{*2} gives more weight to those units that have the smallest variation of the transitory component v_{it} over time. The method of using w_i^{**2} gives more weight to those units that have the largest variation of v_{it} over time. The simple average method gives equal weight to all units. In other words, using w_i^{*2} yields the smoothest evaluation of the estimated ATE over time, while w_i^{**2} yields the most volatile evolution of the ATE over time and using $w_i = \frac{1}{N}$ gives the trend in between. However, if the treatment effects are heterogeneous

$$E(\bar{\Delta}_t) \neq E(\Delta_t^*) \neq E(\Delta_t^{**}), \tag{12.5.12}$$

the projected long-term trend is different. Perhaps a more effective way to inform decision makers could be to provide information on what constitutes the differences in the ATE across individuals at a time or over time.[7]

When there are multiple treatment effects, it also raises the computational issue of whether one should first estimate each individual's treatment effects and then aggregate,

[7] For further discussion on the aggregation issue, see Chapter 14.

or first aggregate all the units that are subject to a treatment at time T_1 and then treat the aggregate unit as a single unit to apply the panel data approach in Section 12.3. Using the mean absolute bias between the estimated ATE(t) and true ATE(t) over time and the mean square error criteria, Hsiao, Shen, and Zhou (2021) show that there is not much difference between the two approaches in their limited Monte Carlo studies. Hence, conditional on the given choice of aggregation method, under the assumption that individual treatment effects can be considered homogeneous, it might be computationally simpler to first aggregate, then apply the methods discussed in Section 12.3.

12.5.2.2 Disaggregate Measure

When treatment effects are heterogeneous, disaggregate measures can provide information on the distribution of the treatment effects. However, it is difficult to provide summary information. Maasoumi and Wang (2019, 2020) suggest using entropy-based measures to summarize the information of the whole distribution of treatment effects: (i) the Kullback-Leibler-Theil information measure,

$$KL = \frac{1}{2}\int \left[\log\left(\frac{f_1}{f_0}\right)\cdot f_1 + \log\left(\frac{f_0}{f_1}\right) f_0\right] dy, \tag{12.5.13}$$

or (ii) the normalized Bhattcharya-Matusita-Hellinger measure (Granger et al. 2004),

$$S = \frac{1}{2}\int \left(f_0^{1/2} - f_1^{1/2}\right)^2 dy \tag{12.5.14}$$

to summarize the distance between two whole distributions where $f_1(y)$ and $f_0(y)$ denote the distribution of y^{1*} and y^{0*}, respectively.

12.6 SIMULATING OR RANKING PROGRAM OUTCOMES WITH DIFFERENT POLICY OPTIONS

12.6.1 Simulating Program Outcomes under Different Policy Scenarios

The discussed panel data approaches to measure the treatment effects, although reasonably simple to implement, are difficult to simulate outcomes under different policy scenarios because the counterfactuals are constructed from observed outcomes of the control units. Observed data are not subject to manipulation. However, under the assumption that policy change does not change the decision rules (i.e., Lucas's (1976) critique does not apply)[8] one may consider simulate outcomes under different policy scenarios through the following steps:

Step 1: Contruct a theoretical model for the outcomes of interest.

Step 2: Estimate the parameters of the theoretical model from observed data.

Step 3: Simulate the potential outcomes under different scenarios.

For instance, Pesaran and Yang (2020) use a stochastic model of epidemics (SIR) on networks to consider COVID-19 infection rate outcome under different policy scenarios. The SIR epidemic model introduced by Kermack and McKendrick (1927) assume

[8] See Damronplasit and Hsiao (2021) for a discussion on policy changes, preference parameters, and dynamic dependence.

1. At any time t, an individual is susceptible (S_t), infected and infectious (I_t), or recovered and immune (R_t)
2. Only susceptible individuals can get infected, remain infectious for some time, and recover and become completely immune.
3. The community is closed. There are no births, deaths, immigration, or emigration during the study period. (Compartment).

Pesaran and Yang (2020) extend the traditional SIR model to a heterogeneous population. The multigroup SIR model further partitions each compartment into multiple groups (e.g., Thieme 2013) by assuming the population for each "compartment" is categorized into L groups of size n_l, $l = 1, 2, \ldots, L$, such that $\frac{n_l}{N} \to w_l \neq 0$ as $N \to \infty$, where $N = \sum_{l=1}^{L} n_l$. Let $d_{il,t}$ be the outcome variable for individual i in the l group at close of day t, with $d_{il,t} = 1$ if "infected" and zero otherwise. The dummy variable $y_{it,l} = 1$ if the infected individual has recovered or died, and 0 otherwise. An individual at time t is considered "active" if he/she is infected and not yet recovered. An "active" individual is represented by

$$z_{il,t} = (1 - y_{il,t}) d_{il,t}. \tag{12.6.1}$$

An individual is "susceptible" if

$$s_{il,t} = 1 - z_{il,t} - y_{il,t} \tag{12.6.2}$$

equals 1, and zero otherwise.

The total number of those "infected" in group l in day t is

$$C_{lt} = \sum_{i=1}^{n_l} d_{il,t}, \quad l = 1, \ldots, L. \tag{12.6.3}$$

The total number of "recovered" or "dead" is

$$R_{lt} = \sum_{i=1}^{n_l} y_{il,t}, \quad l = 1, \ldots, L. \tag{12.6.4}$$

The total number of "active" cases is

$$A_{lt} = C_{lt} - R_{lt}, \quad l = 1, \ldots, L. \tag{12.6.5}$$

The total number of "susceptible" is

$$S_{lt} = n_l - A_{lt} - R_{lt}, \quad l = 1, \ldots, L. \tag{12.6.6}$$

The classic multigroup SIR model can be written as

$$S_{l,t+1} - S_{lt} = -S_{lt} \left(\sum_{l'=1}^{L} b_{ll'} A_{l't} \right), \tag{12.6.7}$$

$$A_{l,t+1} - A_{lt} = S_{lt} \sum_{l'=1}^{L} b_{ll'} A_{l't} - \gamma_l A_{lt}, \tag{12.6.8}$$

$$R_{l,t+1} - R_{lt} = \gamma_l A_{lt}, \tag{12.6.9}$$

where γ_l is the "recovery rate" which is assumed to be time-invariant and the same for all people in group l, and $b_{ll'}$ is the "transmission coefficient" between S_{lt} and $A_{l't}$. Pesaran and Yang (2020) propose to model $d_{il,t}$ as latent variables passing the threshold (e.g., see Chapter 6),[9]

$$d_{il,t+1} = d_{il,t} + (1 - d_{il,t})\ 1(d^*_{il,t+1} > 0), \tag{12.6.10}$$

where $1(A)$ is the indicator function that takes the value of one if A occurs and 0 otherwise. They assume

$$d^*_{il,t+1} = \tau_l \sum_{l'=1}^{L} \sum_{j=1}^{n'_l} p_{il,jl'}(t) z_{il',t} - \xi_{il,t+1}, \tag{12.6.11}$$

where $p_{il,jl'}(t) = 1$ indicates the ith individual in the lth group establishing contacts with the jth individual in the l' group, and 0 otherwise, τ_l denotes the exposure intensity, and $\xi_{il,t+1}$ is randomly distributed over i, l, and t with $E(\xi_{il,t+1}) = \mu_l$ and

$$\text{Prob}(\xi_{il,t+1} < a) = 1 - \exp(-\mu_l^{-1} a),\ a \geq 0. \tag{12.6.12}$$

The recovery process is assumed to follow

$$y_{il,t+1} = y_{il,t} + z_{il,t}\ \eta_{il,t+1}(t^*_{il}), \tag{12.6.13}$$

where $\eta_{il,t+1}(t^*_{il}) = 1$ if (i,l) recovers at time $t+1$ while being infected at t^*_{il} and not before, and 0 otherwise. Under the assumption that the time of removal, $T^*_{il,t} = t - t^*_{il}$, follows a geometric distribution,

$$\text{Prob}(y_{il,t} = 1) = \gamma_l (1 - \gamma_l)^{T^*_{it} - 1},\ T^*_{il,t} = 1, 2, \ldots, \tag{12.6.14}$$

then

$$E(y_{il,t+1} | y_{il,t}, z_{il,t}) = y_{il,t} + \gamma_l z_{il,t}. \tag{12.6.15}$$

Let $p_{ll'} = E(p_{il,jl'}(t))$ and $b_{ll'} = \tau_l k_{ll'}$, where $k_{ll'}$ denotes the average number of contacts per day between individuals in groups l and l'. Aggregating the micro infection moment conditions over i for a group l yields

$$E(C_{l,t+1} - C_{lt} | A^t) = (n_{lt} - C_{lt}) \left[1 - \prod_{l'=1}^{L} (1 - p_{ll'} + p_{ll'} e^{\tau_l})^{A_{l't}} \right], \text{for } l = 1, \ldots, L, \tag{12.6.16}$$

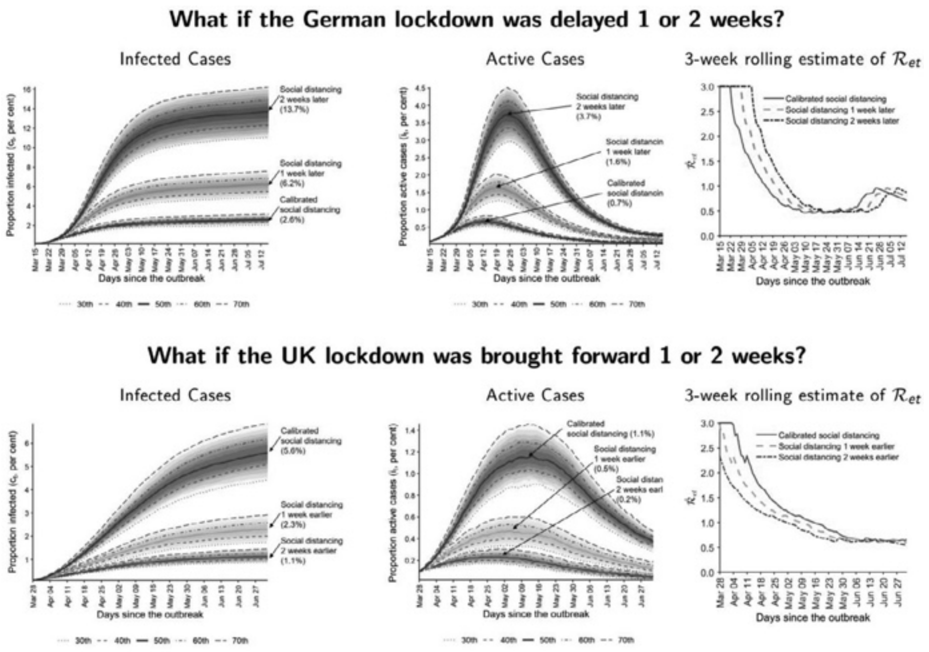
where $A^t = (A_{1t}, \ldots, A_{Lt})'$, and

$$E(R_{l,t+1} | R_{lt}, C_{lt}) = (1 - \gamma_l) R_{lt} + \gamma_l C_{lt}, \quad \text{for } l = 1, \ldots, L. \tag{12.6.17}$$

They use the above moment conditions to estimate the structural parameters, γ_l, τ_l, and $p_{ll'} = p_{l'l}$.

The exposure intensity, τ_l, and the probability of contacts, $p_{ll'}$, can be influenced by policies such as wearing masks and closing of gyms, restaurants, etc.. Pesaran and Yang (2020) simulate different infection rates under different policy scenarios. Figure 12.7 reproduces their figures on the "infected cases" and "active cases" if the German lockdown was delayed one or two weeks and if the UK lockdown was brought forward one or two weeks.

[9] Pesaran and Yang (2021) assume those who "recover" or "die" follow the same process.



Notes: In this exercise recovery is estimated by $R_t = (1 - 1/14)\, R_{t-1} + (1/14) C_{t-1}$ for both countries. Assuming 0.1% infected in day 1 and power law network with mean contact number $k = 10$. $n = 10{,}000$. Number of replications $= 1{,}000$.

Figure 12.7. Counterfactual number of infected and active cases for Germany and UK. Counterfactual number of infected and active cases for Germany and UK.
Note: The simulation uses a single group model with the Erdős–Rényi random network and begins with 1/1,000 of the population randomly infected in day 1; $n = 50{,}000$, $k = 10$, $\gamma = 1/14$. The number of removed (recoveries + deaths) is estimated recursively using $\tilde{R}_t = (1 - \gamma)\, \tilde{R}_{t-1} + \gamma \tilde{C}_{t-1}$ for both countries, with $\tilde{C}_1 = \tilde{R}_1 = 0$. $\hat{\beta}_t$ is the twice-weekly rolling estimate computed with MF $= 5$. The mean of $\hat{\mathcal{R}}_{et}^{(b)} = \left(1 - c_t^{(b)}\right)\hat{\beta}_t/\gamma$ over 1,000 replications is displayed in the last column.
Source: Pesaran and Yang (2020)

12.6.2 Ranking Different Policy Options

When using the moment measures (say ATE), it is usually possible to rank the preference of different policy options. When using disaggregate measurements, it could be difficult to rank different policy options. For instance, consider the measurement of treatment effects between policy option 1, where women and men are paid on the same scale conditional on human capital characteristics (structural effect), and policy option 2, where the current pay scale between men and women remains the same, but women's human capital characteristics become the same as men's (composition effect). For some quantiles of the distributions of treatment effects between options 1 and 2, option 1 could be preferred; but for other quantiles, option 2 could be preferred, as shown by Maasoumi and Wang (2020).

To obtain a unique ranking, Maasoumi and Wang (2019, 2020) suggest a stochastic dominance ranking criterion within the class of weakly increasing utility function $u(y)$:

Suppose $F(y)$ and $G(y)$ are the distribution of treatment effects for options 1 and 2. Let $u(y)$ be every weakly increasing utility function of y; then $F(y)$ is first-order stochastically dominating $G(y)$ if and only if

$$\int u(y)\, dF(y) \geq \int u(y)\, dG(y). \tag{12.6.18}$$

Based on the criterion of (12.6.18), tests of stochastic dominance (e.g., Linton, Maasoumi, and Whang 2005) can be applied for robustness of treatment effects comparison. For instance, consider the policy options about closing the earning gender gap between men and women. Maasoumi and Wang (2019, 2020), using the U.S. Current Population Survey data from 1976 to 2013, constructed the women's counterfactual earning distributions assuming women with the same human capital characteristics as men are rewarded the same in the labor market, the "structural effect"; and women's human capital characteristics are the same as men's, but holding women's wage structure unchanged, the "composition effect." They found that the policy targeting the pay structure is preferred to the policy aimed at changing the human characteristic of women.[10] Their stochastic dominance testing results also suggest that women's human capital characteristics are not necessarily inferior to those of men. Thus, policies aimed at changing only the human characteristics may not produce relative improvements for women.

[10] Maasoumi and Wang (2017, 2020) also applied an inverse probability weighted approach to correct the selection bias in the sample (Firpo et al., 2007).