Econometrics **July 21, 2024**

Topic 6: DID and TWFE

by Sai Zhang

Key points: This note is on the causal panel data, building upon Arkhangelsky and Imbens (2023).

Disclaimer: *This note is compiled by Sai Zhang.*

6.1 Panel Data Configurations

6.1.1 Data Types

6.1.1.1 Panel Data

For observations on N units, indexed by $i = 1, \cdots, N$, over T periods, indexed by $t = 1, \cdots, T$, the outcome of interest is denoted by Y_{it}, the treatment W_{it}. These observations may themselves consist of averages over more basic units:

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & \cdots & Y_{1T} \\ \vdots & \ddots & \vdots \\ Y_{N1} & \cdots & Y_{NT} \end{pmatrix} \qquad \mathbf{W} = \begin{pmatrix} W_{11} & \cdots & W_{1T} \\ \vdots & \ddots & \vdots \\ W_{N1} & \cdots & W_{NT} \end{pmatrix}$$

we may also observe exogenous variables X_{it} or X_i. Typically, we focus on a balanced panel where for all units $i = 1, \cdots, N$ we observe outcomes for all $t = 1, \cdots, T$.

6.1.1.2 Grouped Repeated Cross-Section Data

In a GRCS data, we have observations on N units, each observed only once in period T_i for unit i. Different units may be observed at diffrent points in time, T_i typically takes on only a few values, with many units sharing the same value for T_i. The outcome Y_i and treatment W_i are indexed by the unit index i. The set of units is **partitioned** into 2 or more groups, with the group that unit i belongs to denoted by $G_i \in \mathcal{G} = \{1, 2, \cdots, G\}$.

Define the average outcomes for each group-time-period pair:

$$\overline{Y}_{gt} \equiv \frac{\sum_{i=1}^{N} \mathbf{1}_{G_i=g, T_i=t} Y_i}{\sum_{i=1}^{N} \mathbf{1}_{G_i=g, T_i=t}}$$

for treatment

$$\overline{W}_{gt} \equiv \frac{\sum_{i=1}^{N} \mathbf{1}_{G_i=g, T_i=t} W_i}{\sum_{i=1}^{N} \mathbf{1}_{G_i=g, T_i=t}}$$

then treat the $G \times T$ group averages $\overline{Y}_{gt}$ and $\overline{W}_{gt}$ as the unit of observation, then the grouped data is just a panel. The major issue in practice is that the number of groups is very small comparing to proper panel data.

6.1.1.3 Row and Column Exchangeable Data

The data are doubly indexed by $i = 1, \cdots, N$ and $j = 1, \cdots, J$, with outcomes Y_{ij}. They are different from panel data in that there is **no time ordering** for the second index. Many methods developed for panel data are also applicable here.

6.1.2 Shapes of Data Frames

Panel data can also be loosely classified by the shape:

- **Thin Frames** ($N \gg T$), where the number of cross-section units is large relative to the number of time periods:
  - unit-specific parameters (individual FEs) **can not be estimated consistently** due to the short time series
  - REs might be more suitable since they place a stocahstic structure on the individual components
- **Fat Frames** ($N \ll T$), where the number of cross-section units is large relative to the number of time periods.
- **Square** $N \simeq T$, where the number of units and time periods is comparable.

6.1.3 Assignment Mechanisms

6.1.3.1 The General Case

In the most general case, the treatment may vary both across units and over time, with units **switching** in and out of the treatment group:

$$\mathbf{W}^{\text{general}} = \begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$$

This is more relevant for the RCED configurations, and for panel data of products and promotions as treatments. The assumption on the absence/presence of **dynamic treatment** effects is very important.

6.1.3.2 Single Treated Period

One special case arises when a substantial number of units is treated, but these units are only treated **in the last period**

$$\mathbf{W}^{\text{last}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

If T is relatively small, this case is often analyzed as a cross-section problem, the lagged outcomes are used as exogenous covariates or pre-treatment variables to be adjusted. Here, dynamic effects are not testable,

nor do they matter since the shortness of the panel.

$$\mathbf{W}^{\text{last}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & 1 & \cdots & 0 \end{pmatrix}$$

this setting is prominent in the original applications of the synthetic control literature, here T is usually small.

6.1.3.3 Single Treated Unit and Single Treated Period

An extreme case is where only a single unit is treated, and it is only treated in a single period (typically the last).

$$\mathbf{W}^{\text{block}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Normally, we focus on the effect for the single treated/time-period pair and construct prediction intervals.

6.1.3.4 Block Assignment

The case of block assignment is where a subset of units is treated every period after a common starting date:

$$\mathbf{W}^{\text{block}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & 1 & \cdots & 1 \end{pmatrix}$$

There is typically a sufficient number of treated unit/time-period pairs to allow for reasonable approximations. The presence of dynamic effects change the interpretation of the average effect of the treated: the average effect for the treated now is an average over short **and** medium term effects during different periods.

Staggered Adoption (a.k.a. absorbing treatment setting)

The staggered adoption is the case where units adopt the treatment at various period, and remain in the treatment group once they adopt the treatment:

$$\mathbf{W}^{\text{block}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & 1 & \cdots & 1 \end{pmatrix}$$

Here, with some assumptions, we can separate dynamic effects from heterogeneity across calendar time.

6.1.3.5 Event Study Designs

In the event-study design, units are exposed to the treatment in at most one period:

$$\mathbf{W}^{\text{block}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 & \cdots & 0 \end{pmatrix}$$

There are often dynamic effects of the treatment past hte time of initial treatment, however, the effects might be changing over time.

6.1.3.6 Clustered Assignment

In many applications, units are grouped together in clusters. Units within the same clusters are always assigned to the treatment:

$$\mathbf{W}^{\text{cluster}} = \begin{pmatrix} & & & & & & & \text{cluster} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 3 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & \cdots & 1 & 1 & C \\ 0 & 0 & 0 & 1 & \cdots & 1 & 1 & C \end{pmatrix}$$

Clustering creates complications for inference.

6.1.4 Outcomes, Assumptions and Estimands

For a treatment assignment matrix $\mathbf{W}$, denote:

- the full $T-$component column vector of treatment assignments as

$$\underline{\mathbf{w}} \equiv (w_1, \cdots, w_T)'$$

- the t-component column vector of treatment assignments **up to** time t as

$$\underline{\mathbf{w}}^t \equiv (w_1, \cdots, w_t)'$$

  hence $\underline{\mathbf{w}}^T = \underline{\mathbf{w}}$
- the row vector of treatment values for unit i as $\underline{\mathbf{W}}_i$

Then in general, we can index the potential outcomes for unit i in period t by the full $T-$component vector of assignments $\underline{\mathbf{w}}$

$$Y_{it}\left(\underline{\mathbf{w}}\right)$$

A key underlying assumption is the **Stable Unit Treatment Value Assumption (SUTVA)**, which requires that there is no interference or spillovers between units[1].

[1] SUTVA can hold on a cluster/group level, where the spillover effects are within clusters/groups.

In this setup, there are 2^T potential outcomes for each unit and each time period, as a function of multi-valued treatment $\underline{\mathbf{w}}$. Then, define for each t-unit treatment effects for each pair of assignment vectors $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}'$:

$$\tau_{it}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'} \equiv Y_{it}\left(\underline{\mathbf{w}}'\right) - Y_{it}\left(\underline{\mathbf{w}}\right)$$

and the corresponding population average effect

$$\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'} \equiv \mathbb{E}\left[Y_{it}\left(\underline{\mathbf{w}}'\right) - Y_{it}\left(\underline{\mathbf{w}}\right)\right]$$

where the expectation is implicitly assumed to be taken over a **large** population.

Under completely random assignment, all $\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'}$ are identified, and are **just-identified**, given sufficient variation in the treatment paths. Dynamic treatment effects can also be identified[2]. However, we have $2^{T-1} \times \left(2^T - 1\right)$ distinct average effects of the form $\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'}$, in practice, we often need to focus on summary measures of these causal effects, which requires some addition assumptions:

Assumption 6.1.1: No Anticipation

The potential outcomes satisfy

$$Y_{it}\left(\underline{\mathbf{w}}\right) = Y_{it}\left(\underline{\mathbf{w}}'\right)$$

for all i, and for all combinations of t, $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}'$ such that $\underline{\mathbf{w}}t = \underline{\mathbf{w}}'^t$.

This is a testable assumption with experimental data and sufficient variation in treatment paths, by comparing units that have the same treatment path up to and including t and diverge after t.

- Units **are not** active decision-makers: the assumption can be guaranteed by design (random treatment assignment each period, or staggered adoption with randomly assigned adoption date)
- **Limited** anticipation: assuming the treatment can be anticipated for a **fixed number** of periods, which shifts $\underline{\mathbf{w}}$ by that number of periods.
- Units **are** active decision-makers: potential outcomes are functions of $\underline{\mathbf{w}}$ and the distribution of $\underline{\mathbf{w}}$ (experimental design itself):
  - one can define potential outcomes for a given randomized experimental design: the beliefs about the future treatment paths are incorporated in the definition of the potential outcomes, the actual values are by construction unknown. This does change the interpretation of the casual effects[3].
  - In obserational studies, one cannot directly control the information about the future treatment paths. In this case, different units need to be gauranteed to face the **same information environment** for Assumption 6.1.1 to hold.

Under Assumption 6.1.1, the total number of potential treatment effects is reduced from $2^{T-1} \times \left(2^T - 1\right)$ to $\left(\sum_{t=1}^{T} 2^{t-1}\right)\left(\sum_{t=1}^{T} 2^t - 1\right)$. The unit-period specific treatment effects are now of the type

$$\tau_{it}^{\underline{\mathbf{w}}^t,\underline{\mathbf{w}}^{t'}} \equiv Y_{it}\left(\underline{\mathbf{w}}^{t'}\right) - Y_{it}\left(\underline{\mathbf{w}}^t\right)$$

with the potential outcomes for period t indexed by treatments up to period t only. Here, one can still distinguish

[2]For example, consider that in the 2-period case

$$\tau_{2}^{(1,1),(0,1)}$$

is the average effect in the second period of being exposed to $(1,1)$, *treated in both period*, rather than $(0,1)$, *treated only in the second period*.

[3]Think about the differences between a surprise deviation from a given policy rule versus the effect of a permanent chagne int he policy rule itself.

- static treatment effects: $\tau_{it}^{(\underline{\mathbf{w}}^{t-1},0),(\underline{\mathbf{w}}^{t-1},1)}$, which measures the response of current outcome to the current treatment, holding the past ones fixed.
- dynamic treatment effects: $\tau_{it}^{(\underline{\mathbf{w}}^{t-1},w^t),(\underline{\mathbf{w}}^{t-1},w^t)}$, which does the opposite.

Assumption 6.1.2: No Dynamic/Carry-Over Effects

The potential outcomes satisfy

$$Y_{it}\left(\underline{\mathbf{w}}\right) = Y_{it}\left(\underline{\mathbf{w}}'\right)$$

for all i and for all combinations of t, $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}'$ such that $w_{it} = w'_{it}$.

This assumption is **not** guaranteed by randomization. It restricts the treatment effects and the potential outcomes for the **post**-treatment periods. It has testable restrictions given the random assignment of the treatment and sufficient variation in the treatment paths. It does **not** restrict the time path of the potential outcomes in the absence of any treatment $Y_{it}(\mathbf{0})$.

This assumption greatly reduce the total number of treatment effects for each unit to T:

$$\tau_{it} \equiv Y_{it}(1) - Y_{it}(0)$$

where τ_{it} has no superscripts because there are only 2 possible arguments of the potential outcomes $w \in \{0, 1\}$.

Assumption 6.1.3: Staggered Adoption

In staggered adoption,

$$W_{it} \leq W_{it-1},\ \forall t = 2, \cdots, T$$

define the adoption date A_i as the date of the first treatment, $A_i \equiv T + 1 - \sum_{t=1}^{T} W_{it}$ for treated units, and $A_i \equiv \infty$ for never-treated units.

Under Assumption 6.1.3, the potential outcomes can be written in terms of the adoption date as $Y_{it}(a)$, for $a = 1, \cdots, T, \infty$, and the realized outcome as $Y_{it} = Y_{it}(A_i)$. There are 2 broad classes of settings that are viewed as staggered adoption designs:

- interventions adopted and remain in place
- one-time interventions with a long-term, or even permanent, impact (where the post-intervention period effects are dynamic effects)

Under Assumption 6.1.3, but **not** Assumption 6.1.1 and 6.1.2, we can write

$$\tau_{it}^{a,a'} \equiv Y_{it}(a') - Y_{it}(a)$$

with the corresponding population average

$$\tau_{t}^{a,a'} \equiv \mathbb{E}\left[\equiv Y_{it}(a') - Y_{it}(a)\right]$$

we can also denote the average for subpopulations conditional on the adoption dates as

$$\tau_{t|a''}^{a,a'} \equiv \mathbb{E}\left[\equiv Y_{it}(a') - Y_{it}(a) \mid A_i = a''\right]$$

which explicitly depends on the details of the assignment process. This estimand is conceptually similar to the average effect on the treated in cross-sectional settings, but with selection operating over both unit and period dimensions.

6.1.5 Conventional TWFE and DiD

6.1.5.1 TWFE Characterization

First, consider a panel setting with no anticipation, no dynamics, and constant treatment effects:

Assumption 6.1.4: The TWFE Model

The control outcome $Y_{it}(0)$ satisfies

$$Y_{it}(0) = \alpha_i + \beta_t + \epsilon_{it}$$

The unobserved component ϵ_{it} is (mean-)independent of the treatment assignment W_{it}

And

Assumption 6.1.5: Constant Static Treatment Effects

The potential outcomes satisfy

$$Y_{it}(1) = Y_{it}(0) + \tau \quad \forall (i,t)$$

Under Assumption 6.1.4 and 6.1.5, for the realized $Y_{it} \equiv W_{it} Y_{it}(1) + (1 - W_{it}) Y_{it}(0)$ we have a model

$$Y_{it} = \alpha_i + \beta_t + \tau W_{it} + \epsilon_{it}$$

then we can estimate the parameters of this model by least squares

$$\left(\hat{\tau}^{TWFE}, \hat{\alpha}, \hat{\beta}\right) = \arg\min_{\tau,\alpha,\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \alpha_i - \beta_t - \tau W_{it}\right)^2$$

one restriction on the α_i or β_t needs to be imposed to avoid perfect collinearity, but this normalization does not affect the estimation of τ.

Under a block assignment structure, we have $W_{it} = 1$ only for a subset of the units[4], and those units are treated only during periods t with $t > T_0$. Define the averages in the four groups as

$$\overline{Y}^{\text{tr,post}} \equiv \frac{\sum_{i \in \mathcal{J}} \sum_{t > T_0} Y_{it}}{N^{\text{tr}} (T - T_0)} \qquad \overline{Y}^{\text{tr,pre}} \equiv \frac{\sum_{i \in \mathcal{J}} \sum_{t \leq T_0} Y_{it}}{N^{\text{tr}} T_0}$$

$$\overline{Y}^{\text{co,post}} \equiv \frac{\sum_{i \notin \mathcal{J}} \sum_{t > T_0} Y_{it}}{N^{\text{co}} (T - T_0)} \qquad \overline{Y}^{\text{co,pre}} \equiv \frac{\sum_{i \notin \mathcal{J}} \sum_{t \leq T_0} Y_{it}}{N^{\text{co}} T_0}$$

and then write the estimator for the treatment effect as

$$\hat{\tau}^{TWFE} = \left(\overline{Y}^{\text{tr,post}} - \overline{Y}^{\text{tr,pre}}\right) - \left(\overline{Y}^{\text{co,post}} - \overline{Y}^{\text{co,pre}}\right)$$

6.1.5.2 DiD Estimator in the Grouped Repeated Cross-Section Setting

In GRCS setting, we observe each physical unit only once. With blocked assignment, the notation only has a single index for the unit $i = 1, \cdots, N$. Let $G_i \in \mathcal{G} = \{1, \cdots, G\}$ denote the cluster or group unit i belongs to, and $T_i \in \{1, \cdots, N\}$ the time period unit i is observed in.

[4] The *treatment group* with $i \in \mathcal{J}$, where the cardinality for the set $\mathcal{J}$ is N^{tr} and $N^{\text{co}} \equiv N - N^{\text{tr}}$.

The set of clusters $\mathcal{G}$ is partitioned into two groups: control group $\mathcal{G}_C$ and treatment group $\mathcal{G}_T$, with cardinality G_C and G_T. Only units with $G_i \in \mathcal{G}_T$, indicated by $D_i = \mathbf{1}_{G_i \in \mathcal{G}_T}$, are exposed to the treatment if they are observed after the treatment date T_0: $W_i = \mathbf{1}_{G_i \in \mathcal{G}_T, T_i > T_0}$

Assuming that the treatment within group and time period pairs is constant, the cluster/time-period average treatment $\overline{W}_{gt}$ is binary if the original treatment is. Then the DiD estimator is

$$\begin{aligned}\hat{\tau}^{DiD} =& \frac{1}{G_T\,(T - T_0)} \sum_{g \in \mathcal{G}_T, t > T_0} \overline{Y}_{gt} - \frac{1}{G_C\,(T - T_0)} \sum_{g \in \mathcal{G}_C, t > T_0} \overline{Y}_{gt} \\ & - \frac{1}{G_T T_0} \sum_{g \in \mathcal{G}_T, t \le T_0} \overline{Y}_{gt} + \frac{1}{G_C T_0} \sum_{g \in \mathcal{G}_C, t \le T_0} \overline{Y}_{gt}\end{aligned}$$

and at the group level, we have a proper panel setup:

$$\overline{Y}_{gt}(0) = \alpha_g + \beta_t + \epsilon_{gt} \qquad\qquad \overline{Y}_{gt}(1)\overline{Y}_{gt}(0) + \tau$$

and the potential outcomes $\overline{Y}_{gt}(0)$ and $\overline{Y}_{gt}(1)$ should be interpreted as the average of the potential outcomes if all units in a group/time-period pair are exposed to the control treatment.

6.1.5.3 Inference

There are two ways to conduct inference about $\hat{\tau}^{\text{DiD}}$ and $\hat{\tau}^{\text{TWFE}}$:

- the assignment process is known: **design-based** or **randomization-based** inference
- otherwise: **sampling-based** inference

Design-Based Inference

Sampling-Based Inference

- **proper panel setting**: it is often assumed that all units are randomly sampled from a large population and thus **exchangeable**. Inference about $\hat{\tau}^{\text{TWFE}}$ reduces to joint inference about four means with i.i.d. observations.
- **GRCS setting**: one can allow for non-vanishing errors at the group level, but it cannot be done in the two-group case.

Standard errors Regardless of the level of aggregation, inference for TWFE and DiD estimators typically takes into account the correlation in outcomes over time within units in applications with more than two periods. So it is **NOT** appropriate to use the robust Eicker-Huber-White standard errors. Instead, one should use clustered standard errors based on clustering observations by units. It can also be approximated by bootstrapping all observations for each unit.

6.1.5.4 The Parallel Trend Assumption

The **parallel trend assumption** is the fundamental justification for the DiD estimator. It states that the units who are treated would have followed a path that is parallel to the path followed by the control units on average, in the absence of the treatment.

Proper panel settings the assumption is that the expected difference in control outcomes in any period for units who later are exposed to the treatment and units who are always in the control group is **contstant**:

Assumption 6.1.6: Parallel Trend Assumption: Proper Panel

For all t, t',

$$\mathbb{E}\left[Y_{it}\left(0\right) \mid D_i = 1\right] - \mathbb{E}\left[Y_{it}\left(0\right) \mid D_i = 0\right] = \mathbb{E}\left[Y_{it'}\left(0\right) \mid D_i = 1\right] - \mathbb{E}\left[Y_{it'}\left(0\right) \mid D_i = 0\right]$$

equivalently, we can formulate it in terms of changes over time[a]:

$$\mathbb{E}\left[Y_{it}(0) - Y_{it'}(0) \mid D_i = 1\right] = \mathbb{E}\left[Y_{it}(0) - Y_{it'}(0) \mid D_i = 0\right]$$

alternatively, postulate a TWFE model for the control outcomes, additionally assuming that the treatment assignment D_i is independent of the vector of residuals $\epsilon_{it}, t = 1, \cdots, T$, conditional on FEs:

$$D_i \perp \left(\epsilon_{i1}, \cdots, \epsilon_{iT}\right) \mid \alpha_i$$

[a] the expected change in control outcomes is the same for those who will eventually be exposed to the treatment and those who will not

From the point of view of the modern casual inference literature, the parallel trend assumption is non-standard in the sense that it combbines restrictions on the potential outcomes with restrictions on the assignment mechanism.

GRCS settings Suppose in the population, all groups are (infinitely) large in each period, and we have random samples from these populations for each period. Then the expectations are well defined as population averages. The parallel trends assumption can the nbe formulated as requiring that the difference in expected control outcomes between two groups remains constant over time:

Assumption 6.1.7: Parallel Trend Assumption: Grouped Repeated Cross-Section

For all pairs of groups g, g' and for all pairs of time periods t, t', the average difference between the groups remains the same over time, irrespective of their treatment status:

$$\mathbb{E}\left[Y_{gt}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{g't}(0) \mid D_i = 0\right] = \mathbb{E}\left[Y_{gt'}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{g't'}(0) \mid D_i = 0\right]$$

an alternative formulation is that expected change between periods t' and t is the same for all groups:

$$\mathbb{E}\left[Y_{gt}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{gt'}(0) \mid D_i = 1\right] = \mathbb{E}\left[Y_{g't}(0) \mid D_i = 0\right] - \mathbb{E}\left[Y_{g't'}(0) \mid D_i = 0\right]$$

If $Y_{gt}(0)$ for all g and t are observed, the presence of the two groups and two time periods would be sufficient for the assumption to have testable implications. However, in the 2-group/2-period case, at least one of the four cells is exposed to the treatment, there are no testable restrictions implied by this assumption[5].

6.1.5.5 Pre-treatment Variables

Time-invariant characteristics of the units in addition to the time path of the outcome are observed, these variables are colinear with the individual fixed effects α_i hence cannot be incorporated simply by adding them to the TWFE specification. A reason one might want to include these pre-treatment variables is that

[5] If there are more than 2 periods, or more than 2 groups, there are testable restrictions by the parallel trend assumption.

the parallel trend and constant treatment effect assumptions hold only within subpopulations defined by them.

Semi-parametric DiD Abadie (2005) proposed a solution based on **re-weighting** the differences in outcomes by the propensity score for balance. TO estimate the average treatment effect on the treated (ATT):

$$ATT \equiv \mathbb{E}\left(\mathbf{y}_{1t} - \mathbf{y}_{0t} \mid \mathbf{d} = 1\right)$$

where the 2 potential outcomes $\mathbf{y}_{1t}$ is the value of $\mathbf{y}$ if the participant received the treatment by t, $\mathbf{y}_{0t}$ is the value of $\mathbf{y}$ if the participant had not received the treatment by time t. $\mathbf{d}$ is an indicator of treatment.

ATT cannot be directly estimated since $\mathbf{y}_{0t}$, the counterfactual, is never observed. For a set of pretreatment characteristics $\mathbf{x}_b$, define the probability to be in the treatment group conditional on $\mathbf{x}_b$ as

$$\pi\left(\mathbf{x}_b\right) \equiv \mathbb{P}\left(\mathbf{d} = 1 \mid \mathbf{x}_b\right)$$

define the change of $\mathbf{y}$ from baseline b to t as

$$\Delta\mathbf{y}_t \equiv \mathbf{y}_t - \mathbf{y}_b$$

then

$$\mathbb{E}\left\{\frac{\Delta\mathbf{y}_t}{\mathbb{P}\left(\mathbf{d}=1\right)} \times \frac{\mathbf{d} - \pi\left(\mathbf{x}_b\right)}{1 - \pi\left(\mathbf{x}_b\right)}\right\} \tag{6.1}$$

gives an unbiased estimate of the ATT if

$$\begin{gathered}\mathbb{E}\left(\mathbf{y}_{0t} - \mathbf{y}_{0b} \mid \mathbf{d} = 1, \mathbf{x}_b\right) = \mathbb{E}\left(\mathbf{y}_{0t} - \mathbf{y}_{0b} \mid \mathbf{d} = 0, \mathbf{x}_b\right) \\ \mathbb{P}\left(\mathbf{d}=1\right) > 0 \\ \pi\left(\mathbf{x}_b\right) < 1\end{gathered}$$

This estimator is a weighted average of the difference of trend, $\Delta\mathbf{y}_t$, across treatment groups: it reweights the trend of the untreated based on the propensity score $\pi\left(\mathbf{x}_b\right)$[6].

Abadie (2005) suggests to approximate the propensity score $\pi\left(\mathbf{x}_b\right)$ semiparametrically using a polynomial series of the predictors and plug the predicted values into the sample analogue of the ATT estimates 6.1. There are two main ways to do the approximation:

- **linear probability model (LPM)**: higher order improves the approximation, but less precise
- **series logit estimator (SLE)**: using a logit specification to constrain the estimated propensity score to vary between 0 and 1

consider $\hat{\pi}\left(\mathbf{x}_b\right)$, the approximated propensity score, and k, the order of the polynomial function for approximation. Then the **LPM** approximation is

$$\hat{\pi}\left(\mathbf{x}_b\right) = \hat{\gamma}_0 + \hat{\gamma}_1 \times \mathbf{x}_1 + \sum_{i=1}^{k} \hat{\gamma}_{2i} \times \mathbf{x}_2^i$$

where $\mathbf{x}_1$ is a binary variable, $\mathbf{x}_2^i = \prod_{j=1}^{i} \mathbf{x}_2$, with $\mathbf{x}_2$ being a continuous variable. Then the coefficients $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_{21}, \cdots, \hat{\gamma}_{2i}, \cdots, \hat{\gamma}_{2k}$ are estimated using OLS estimators.

[6] $\frac{\pi(\mathbf{x}_b)}{1-\pi(\mathbf{x}_b)}$ is an increasing function of $\pi\left(\mathbf{x}_b\right)$, hence untreated participants with a higher propensity score are given a higher weight.

The **SLE** approximation is

$$\hat{\pi}\left(\mathbf{x}_b\right)=\Lambda\left(\hat{\gamma}_0+\hat{\gamma}_1\times\mathbf{x}_1+\sum_{k=1}^{K}\hat{\gamma}_{2k}\times\mathbf{x}_2^k\right)$$

where $\Lambda(x)=\frac{\exp(x)}{1+\exp(x)}$ is the logistic function. Higher order binary variables are not considered here since $\mathbf{x}_1^k=\mathbf{x}_1$ for any value $k>1$.

Doubly robust DiD Sant'Anna and Zhao (2020) adjust for time-invariant covariates in a doubly robust way, by combining inverse-propensity score weighting with outcome modeling.

Timing varying covariates With finite T, strictly exogenous time-varying covariates X_{it} can be converted to time invariant $X_i\equiv(X_{i1},\cdots,X_{iT})$, in practice, applied researchers only rely on linear specifications with contemporaneous covariates instead.

Sant'Anna and Zhao (2020) also assume that covariates and treatment status are stationary as Abadie (2005). Let T_i be a dummy variable that takes value one if the observation i is only observed in the post-treatment period, and 0 if only observed in the pre-treatment period. Define $Y_i=T_iY_{i1}+(1-T_i)Y_{i0}$. Let n_1 and n_0 be the sample size of the post- and pre-treatment periods such that $n=n_1+n_0$, and let $\lambda=\mathbb{P}(T=1)\in(0,1)$:

Assumption 6.1.8: Main assumptions of Sant'Anna and Zhao (2020)

Assume that

1 the data $\{Y_{i0},Y_{i1},D_i,X_i\}_{i=1}^n$ are **i.i.d.**, or the pooled repeated cross-section data $\{Y_i,D_i,X_i,T_i\}_{i=1}^n$ consisting of i.i.d. draws from the mixture distribution

$$\begin{aligned}\mathbb{P}\left(Y\le y,D=d,X\le x,T=t\right)=&t\cdot\lambda\cdot\mathbb{P}\left(Y_1\le y,D=d,X\le x\mid T=1\right)\\&+(1-t)\cdot(1-\lambda)\mathbb{P}\left(Y_0\le y,D=d,X\le x\mid T=0\right)\end{aligned}$$

where $\left(y,d,x,t\right)\in\mathbb{R}\times\{0,1\}\times\mathbb{R}^k\times\{0,1\}$, with the joint distribution of (D,X) invariant to T.

2 **Conditional Parallel Trend Assumption (PTA)**[a]:

$$\mathbb{E}\left[Y_1(0)-Y_0(0)\mid D=1,X\right]\overset{a.s.}{=}\mathbb{E}\left[Y_1(0)-Y_0(0)\mid D=1,X\right]$$

3 $\exists\epsilon>0,\mathbb{P}(D=1)>\epsilon$ and $\mathbb{P}(D=1\mid X)\le1-\epsilon$ a.s.[b]

[a] It allows for covariate-specific time trends but not unit specific trends.

[b] This overlapping condition states that at least a small fraction of the population is treated and that for every value of X, at least a small probability that the unit is not treated.

Under Assumption 6.1.8, there are 2 main flexible estimation procedures to estimate the ATT:

1 outcome regression (**OR**) approach

$$\hat{\tau}^{\text{reg}}=\overline{Y}_{1,1}-\left[\overline{Y}_{1,0}+n_{\text{treat}}^{-1}\sum_{i|D_i=1}\left(\hat{\mu}_{0,1}\left(X_i\right)-\hat{\mu}_{0,0}\left(X_i\right)\right)\right]$$

where $\overline{Y}_{d,t}=\sum_{i|D_i=d,T_i=t}Y_{it}/n_{d,t}$ is the sample average outcome among units in treatment group d and time t, $\hat{\mu}_{d,t}(x)$ is an estimator of the unknown $m_{d,t}(x)\equiv\mathbb{E}\left[Y_t\mid D=d,X=x\right]$

2 inverse propensity weighting (**IPW**) approach, as in Abadie (2005).

Sant'Anna and Zhao (2020) proposed to combine both the **OR** and **IPW** approaches to form the doubly robust (**DR**) moments/estimands for the ATT.

Notation Let $\pi(X)$ be an arbitrary model for the true, unknown propensity score.

- with proper panel data, let $\Delta Y = Y_1 - Y_0$ and define $\mu^p_{d,\Delta}(X) \equiv \mu^p_{d,1}(X) - \mu^p_{d,0}(X)$ being a model for the true, unknown outcome regression $m^p_{d,t}(x) \equiv \mathbb{E}[Y_t \mid D = d, X = x]$, $d, t = 0, 1$.
- with repeated cross-section data, let $\mu^{rc}_{d,t}(x)$ be an arbitrary model for the true, unknown regression $m^{rc}_{d,t}(x) \equiv \mathbb{E}[Y \mid D = d, T = t, X = x]$, $d, t = 0, 1$, $\mu^{rc}_{d,Y}(T, X) \equiv T \cdot \mu^{rc}_{d,1}(X) + (1 - T) \cdot \mu^{rc}_{d,0}(X)$, and $\mu^{rc}_{d,\Delta}(X) \equiv \mu^{rc}_{d,1}(X) - \mu^{rc}_{d,0}(X)$.

Estimands consider

- for proper panel data:

$$\tau^{dr,p} = \mathbb{E}\left[\left(w^p_1(D) - w^p_0(D, X; \pi)\right)\left(\Delta Y - \mu^p_{0,\Delta}(X)\right)\right]$$

where, for a generic g,

$$w^p_1(D) = \frac{D}{\mathbb{E}[D]} \qquad w^p_0(D, X; g) = \frac{g(X)(1 - D)}{1 - g(X)} \cdot \left(\mathbb{E}\left[\frac{g(X)(1 - D)}{1 - g(X)}\right]\right)^{-1}$$

- for repeated cross-section data, consider 2 different estimands

$$\begin{aligned}
\tau^{dr,rc}_1 =& \mathbb{E}\left[\left(w^{rc}_1(D, T) - w^{rc}_0(D, T, X; \pi)\right) \cdot \left(Y - \mu^{rc}_{0,Y}(T, X)\right)\right] \\
\tau^{dr,rc}_2 =& \tau^{dr,rc}_1 + \left(\mathbb{E}\left[\mu^{rc}_{1,1}(X) - \mu^{rc}_{0,1}(X) \mid D = 1\right] - \mathbb{E}\left[\mu^{rc}_{1,1}(X) - \mu^{rc}_{0,1}(X) \mid D = 1, T = 1\right]\right) \\
& - \left(\mathbb{E}\left[\mu^{rc}_{1,0}(X) - \mu^{rc}_{0,0}(X) \mid D = 1\right] - \mathbb{E}\left[\mu^{rc}_{1,0}(X) - \mu^{rc}_{0,0}(X) \mid D = 1, T = 0\right]\right)
\end{aligned}$$

where for a generic g,

$$w^{rc}_1(D, T) = w^{rc}_{1,1}(D, T) - w^{rc}_{1,0}(D, T) \qquad w^{rc}_0(D, T, X; g) = w^{rc}_{0,1}(D, T, X; g) - w^{rc}_{0,0}(D, T, X; g)$$

and for $t = 0, 1$

$$\begin{aligned}
w^{rc}_{1,t}(D, T) &= \frac{D \cdot 1\{T = t\}}{\mathbb{E}[D \cdot 1\{T = t\}]} \\
w^{rc}_{0,t}(D, T, X; g) &= \frac{g(X)(1 - D) \cdot 1\{T = t\}}{1 - g(X)} \left(\mathbb{E}\left[\frac{g(X)(1 - D) \cdot 1\{T = t\}}{1 - g(X)}\right]\right)^{-1}
\end{aligned}$$

Then if at least

- for **panel** data, either $\pi(X) \overset{a.s.}{=} p(X)$ or $\mu^p_\Delta(X) \overset{a.s.}{=} m^p_{0,1}(X) - m^p_{0,0}(X)$
- for **repeated cross-section** data, either $\pi(X) \overset{a.s.}{=} p(X)$ or $\mu^p_{0,\Delta}(X) \overset{a.s.}{=} m^{rc}_{0,1}(X) - m^{rc}_{0,0}(X)$[7]

that is, at least one of the working nuisance models is correctly specified, the ATT can be estimated. This is less demanding than both OR and IPW approach.

[7]For repeated cross-section data, $\tau^{dr,rc}_1$ does not reply on OR models for the treated group but $\tau^{dr,rc}_2$ does, however, $\tau^{dr,rc}_1$ is not more robust against model misspecification than $\tau^{dr,rc}_2$ since they identify the ATT under the same conditions. Given that $\mathbb{E}\left[g(X) \mid D = 1\right] = \mathbb{E}\left[g(X) \mid D = 1, T = t\right]$, $t = 0, 1$ holds for any $g(\cdot)$, it must hold for $\mu^{rc}_{1,t}(\cdot) - \mu^{rc}_{0,t}(\cdot)$, $t = 0, 1$, even when $\mu^{rc}_{d,t}(\cdot)$ are misspecified.

Semiparametric efficiency bound Let $m^p_{0,\Delta} \equiv m^p_{0,1}(x) - m^p_{0,0}(x)$ and $m^{rc}_{d,\Delta}(X) \equiv m^{rc}_{0,1}(X) - m^{rc}_{0,0}(X)$ for $d = 0, 1$. Then

- for **panel** data, the **efficient influence function** for the ATT is

$$\begin{aligned}\eta^{e,p}(Y_1, Y_0, D, X) =& w^p_1(D)\left(m^p_{1,\Delta}(X) - m^p_{0,\Delta}(X) - \tau\right) \\ &+ w^p_1(D)\left(\Delta Y - m^p_{1,\Delta}(X)\right) - w^p_0(D, X; p)\left(\Delta Y - m^p_{0,\Delta}(X)\right)\end{aligned}$$

  and the **semiparametric efficiency bound** for all regular ATT estimators is

$$\begin{aligned}\mathbb{E}\left[\eta^{e,p}(Y_1, Y_0, D, X)\right]^2 =& \frac{1}{\mathbb{E}[D]^2}\left[D\left(m^p_{1,\Delta}(X) - m^p_{0,\Delta}(X) - \tau\right)^2\right. \\ &\left. + D\left(\Delta - m^p_{1,\Delta}(X)\right)^2 + \frac{(1-D)p(X)^2}{\left(1-p(X)\right)^2}\left(\Delta Y - m^p_{0,\Delta}(X)\right)^2\right]\end{aligned}$$

- for **repeated cross-section** data, the **efficient influence function** for the ATT is

$$\begin{aligned}\eta^{e,rc}(Y, D, T, X) =& \frac{D}{\mathbb{E}[D]}\left(m^{rc}_{1,\Delta}(X) - m^{rc}_{0,\Delta}(X) - \tau\right) \\ &+ \left(w^{rc}_{1,1}(D, T)\left(Y - m^{rc}_{1,1}(X)\right) - w^{rc}_{1,0}(D, T)\left(Y - m^{rc}_{1,0}(X)\right)\right) \\ &- \left(w^{rc}_{0,1}(D, T, X; p)\left(Y - m^{rc}_{0,1}(X)\right) - w^{rc}_{0,0}(D, T, X; p)\left(Y - m^{rc}_{0,0}(X)\right)\right)\end{aligned}$$

  and the **semiparametric efficiency bound** for all regular ATT estimators is

$$\begin{aligned}\mathbb{E}\left[\eta^{e,rc}(Y, D, T, X)^2\right] =& \frac{1}{\mathbb{E}[D]^2}\mathbb{E}\left[D\left(m^{rc}_{1,\Delta}(X) - m^{rc}_{0,\Delta}(X) - \tau\right)^2\right. \\ &+ \frac{DT}{\lambda^2}\left(Y - m^{rc}_{1,1}(X)\right)^2 + \frac{D(1-T)}{(1-\lambda)^2}\left(Y - m^{rc}_{1,0}(X)\right)^2 \\ &\left. + \frac{(1-D)p(X)^2T}{\left(1-p(X)\right)^2\lambda^2}\left(Y - m^{rc}_{0,1}(X)\right)^2 + \frac{(1-D)p(X)^2(1-T)}{\left(1-p(X)\right)^2(1-\lambda)^2}\left(Y - m^{rc}_{0,0}(X)\right)^2\right]\end{aligned}$$

Both $\eta^{e,p}$ and $\eta^{e,rc}$ depends on the true, unknown, outcome regression functions for the treated group, $m_{1,1}(\cdot)$ and $m_{1,0}(\cdot)$ in an asymmetric manner.

The key difference between the two estimators is that for panel data,

$$\begin{aligned}\eta^{e,p}(Y_1, Y_0, D, X) =& w^p_1(D)\left(m^p_{1,\Delta}(X) - m^p_{0,\Delta}(X) - \tau\right) \\ &+ w^p_1(D)\left(\Delta Y - m^p_{1,\Delta}(X)\right) - w^p_0(D, X; p)\left(\Delta Y - m^p_{0,\Delta}(X)\right) \\ =& \left[w^p_1(D) - w^p_0(D, X; p)\right] Y\left[\Delta Y - m_{0,\Delta}(X)\right] - w^p_1(D)\cdot\tau\end{aligned}$$

which ends up **not** depending are $m_{1,1}(\cdot)$ or $m_{1,0}(\cdot)$.

Comparing the efficiency bound of the two cases, we have, if T is independent of (Y_1, Y_0, D, X),

$$\begin{aligned}
&\mathbb{E}\left[\eta^{e,rc}(Y,D,T,X)^2\right] - \mathbb{E}\left[\eta^{e,p}(Y_1,Y_0,D,X)^2\right] \\
=&\frac{1}{\mathbb{E}[D]^2}\left[D\left(\sqrt{\frac{1-\lambda}{\lambda}}(Y_1 - m_{1,1}(X)) + \sqrt{\frac{\lambda}{1-\lambda}}(Y_0 - m_{1,0}(X))\right)^2\right. \\
&\left. + \frac{(1-D)P(X)^2}{\left(1-p(X)\right)^2}\left(\sqrt{\frac{1-\lambda}{\lambda}}(Y_1 - m_{0,1}(X)) + \sqrt{\frac{\lambda}{1-\lambda}}(Y_0 - m_{0,0}(X))\right)^2\right] \geq 0
\end{aligned}$$

which gives that under the DiD framework, it is possible to form more efficient estimators for the ATT when the panel data are available.

This result also gives an efficiency-loss-minimizing λ:

$$\lambda = \frac{\tilde{\sigma}_1}{\tilde{\sigma}_0 + \tilde{\sigma}_1} \qquad \text{where } \tilde{\sigma}_t^2 = \mathbb{E}\left[D\left(Y_t - m_{1,t}(X)\right)^2 + \frac{(1-D)p(X)^2}{\left(1-p(X)\right)^2}\left(Y_t - m_{0,t}(X)\right)^2\right], t = 0,1$$

hence, in principle, one may benefit from *oversampling* from either the pre- or post-treamtent period. But it's generally infeasible to do so during the design stage since $\tilde{\sigma}_1^2$ depends on post-treatment data. Sant'Anna and Zhao (2020) recommand $\lambda = 0.5$ for DiD with repeated cross-section units as a reasonable choice.

Estimation and inference Sant'Anna and Zhao (2020) proposed a two-step procedure for estimation:

- first, estimate the true, unknown $p(\cdot)$ with $\pi(\cdot)$, the true unknown $m_{d,t}^p(\cdot)$ and $m_{d,t}^{rc}(\cdot)$ with $\mu_{d,t}^p(\cdot)$ and $\mu_{d,t}^{rc}(\cdot)$, $d, t = 0, 1$
- second, plug the fitted values of the estimated propensity score and regression models into the sample analogue of $\tau^{dr,p}$, $\tau_1^{dr,rc}$, $\tau_2^{dr,rc}$

instead of using semi-parametric estimators as Abadie (2005), Sant'Anna and Zhao (2020) use generic parametric estimators for the first step, assuming:

- $\pi\left(x;\gamma^*\right)$ is a parametric model for $p(x)$ s.t. $\pi(\cdot)$ is known up to the **finite** dimensional pseudo-true γ^*
- for $d, t = 0, 1$, $\mu_{d,t}^p\left(x;\beta_{d,t}^{*,p}\right)$ and $\mu_{d,t}^{rc}\left(x;\beta_{d,t}^{*,rc}\right)$ s.t. they are known up to the finite dimensional pseudo-true parameter $\beta_{d,t}^{*,p}$ and $\beta_{d,t}^{*,rc}$

this approach is most suitable when the sample size is moderate and the dimension of covariates is high. The estimations are

- for **panel data**

$$\hat{\tau}^{dr,p} = \mathbb{E}_n\left[\left(\hat{w}_1^p(D) - \hat{w}_0^p\left(D,X;\hat{\gamma}\right)\right)\left(\Delta Y - \mu_{0,\Delta}^p\left(X;\hat{\beta}_{0,0}^p,\hat{\beta}_{0,1}^p\right)\right)\right]$$

  where

$$\hat{w}_1^p(D) = \frac{D}{\mathbb{E}_n[D]} \qquad \hat{w}_0^p\left(D,X;\gamma\right) = \frac{\pi(X;\gamma)(1-D)}{1-\pi(X;\gamma)}\left(\mathbb{E}_n\left[\frac{\pi(X;\gamma)(1-D)}{1-\pi(X;\gamma)}\right]\right)^{-1}$$

- for **repeated cross-section data**,

$$\begin{aligned}\hat{\tau}_1^{dr,rc} =& \mathbb{E}_n\left[\left(\hat{w}_1^{rc}(D,T) - \hat{w}_0^{rc}(D,T,X;\hat{\gamma})\right)\left(Y - \mu_{0,Y}^{rc}\left(T,X;\hat{\beta}_{0,0}^{rc},\hat{\beta}_{0,1}^{rc}\right)\right)\right] \\ \hat{\tau}_2^{dr,rc} =& \hat{\tau}_1^{dr,rc} + \left(\mathbb{E}_n\left[\left(\frac{D}{\mathbb{E}_n[D]} - \hat{w}_{1,1}^{rc}(D,T)\right)\left(\mu_{1,1}^{rc}\left(X;\hat{\beta}_{1,1}^{rc}\right) - \mu_{0,1}^{rc}\left(X;\hat{\beta}_{0,1}^{rc}\right)\right)\right]\right) \\ & - \left(\mathbb{E}_n\left[\left(\frac{D}{\mathbb{E}_n[D]} - \hat{w}_{1,0}^{rc}(D,T)\right)\left(\mu_{1,0}^{rc}\left(X;\hat{\beta}_{1,0}^{rc}\right) - \mu_{0,1}^{rc}\left(X;\hat{\beta}_{0,0}^{rc}\right)\right)\right]\right)\end{aligned}$$

  where

$$\begin{aligned}\mu_{0,Y}^{rc}\left(Y,X;\beta_{0,0}^{rc},\beta_{0,1}^{rc}\right) &= T\cdot\mu_{0,1}^{rc}\left(\cdot;\beta_{0,1}^{rc}\right) + (1-T)\mu_{0,0}^{rc}\cdot\left(\cdot;\beta_{0,0}^{rc}\right) \\ \mu_{d,\Delta}^{rc}\left(\cdot;\beta_{d,1}^{rc},\beta_{d,0}^{rc}\right) &= \mu_{d,1}^{rc}\left(\cdot;\beta_{d,1}^{rc}\right) - \mu_{d,0}^{rc}\left(\cdot;\beta_{d,0}^{rc}\right)\end{aligned}$$

These estimators can be improved to achieve not only **consistency** doubly robustness, but also **inference** doubly robustness[8]. For the improvement, Sant'Anna and Zhao (2020) assume, in addition,

- linear regression working models, for the outcome of interest
- a logistic working model, for the propensity score
- covariates X entering all nuisance models in a symemtric manner

which are more stringent than the generic DR DiD estimators, but weaker than TWFE estimators. Under such assumptions, we have the improved DR DiD estimators

- for **panel data**, the 3-step estimator is given as

$$\hat{\tau}_{imp}^{dr,p} = \mathbb{E}_n\left[\left(\hat{w}_1^p(D) - \hat{w}_0^p\left(D,X;\hat{\gamma}^{ipt}\right)\right)\left(\Delta Y - \mu_{0,\Delta}^{lin,p}\left(X;\hat{\beta}_{0,\Delta}^{wls,p}\right)\right)\right]$$

  the first two steps compute

$$\hat{\gamma}^{ipt} = \arg\max_{\gamma\in\Gamma}\mathbb{E}_n\left[DX'\gamma - (1-D)\exp(X'\gamma)\right], \quad \hat{\beta}_{0,\Delta}^{wls,p}\arg\min_{b\in\Theta}\mathbb{E}_n\left[\frac{\Lambda\left(X'\hat{\gamma}^{ipt}\right)}{1-\Lambda\left(X'\hat{\gamma}^{ipt}\right)}(\Delta Y - X'b)^2 \mid D=0\right]$$

  where $\hat{\gamma}^{ipt}$ is the inverse probability tilting estimator, $\hat{\beta}_{0,\Delta}^{wls,p}$ is the weighted least squares estimator for $\beta_{0,\Delta}^{*,p}$. In the last step, plug the fitted value of the working models for the nuisance functions

$$\pi(X,\gamma) = \Lambda\left(X'\gamma\right) \equiv \frac{\exp(X'\gamma)}{1+\exp(X'\gamma)} \qquad \mu_{0,\Delta}^{p}\left(X;\beta_{0,1}^{p},\beta_{0,1}^{p}\right) = \mu_{0,\Delta}^{lin,p}\left(X;\beta_{0,\Delta}^{p}\right) \equiv X'\beta_{0,\Delta}^{p}$$

  into the sample analogue of $\tau^{dr,p}$. Sant'Anna and Zhao (2020) show that if

$$\mathbb{E}\left[\left(\frac{D}{\mathbb{E}[D]} - \frac{\exp(X'\gamma^*)(1-D)}{\mathbb{E}\left[\exp(X'\gamma^*)(1-D)\right]}\right)X\right] = 0 \qquad \mathbb{E}\left[\exp(X'\gamma^*)\left(\Delta Y - \mu_{0,\Delta}^{lin,p}\left(X;\beta_{0,\Delta}^{*}\right)\right)X \mid D=0\right] = 0$$

  there will be no estimation effect from the first stage, with the linear outcome models and logistic propensity score models assumed.

[8]This way, there is no estimation effect from first-step estimators, the asymptotic variance of the results DR DiD estimator for the ATT is invariant to which working modesl for the nuisance functions are correctly specified. This in practice usually translates to simpler and more stable inference procedures.

As $n \to \infty$, these 2 moment conditions follow from the first-order conditions of the optimization problems associated with $\hat{\gamma}^{ipt}$ and $\hat{\beta}_{0,\Delta}^{wls,p}$, even when the working models are misspecified. Hence, replacing the pseudo-true parameters $\gamma^{*,ipt}$ and $\beta_{0,\Delta}^{*,wls,p}$ with their estimators $\hat{\gamma}^{ipt}$ and $\hat{\beta}_{0,\Delta}^{wls,p}$ guarantee that $\hat{\tau}_{imp}^{dr,p}$ is doubly robust:

$$\hat{\tau}_{imp}^{dr,p} \xrightarrow{p} \tau, \quad \text{if } \textbf{either } \Lambda(X'\gamma^{*,ipt}) \overset{a.s.}{=} p(X) \textbf{ or } X'\beta_{0,\Delta}^{*,wls,p} \overset{a.s.}{=} m_{0,\Delta}^{p}(X)$$

As for inference, $\hat{\tau}_{imp}^{dr,p}$ is $\sqrt{n}-$consistent and asymptotically normal

$$\sqrt{n}\left(\hat{\tau}_{imp}^{dr,p} - \tau_{imp}^{dr,p}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \eta_{imp}^{dr,p}\left(W; \gamma^{*,ipt}, \beta_{0,\Delta}^{*,wls,p}, \tau_{imp}^{dr,p}\right) + o_p(1) \xrightarrow{d} \mathcal{N}\left(0, V_{imp}^{p}\right)$$

and if **both** $\Lambda(X'\gamma^{*,ipt}) \overset{a.s.}{=} p(X)$ **and** $X'\beta_{0,\Delta}^{*,wls,p} \overset{a.s.}{=} m_{0,\Delta}^{p}(X)$, V_{imp}^{p} equals to the semiparametrically efficiency bound, and it can be estimated as

$$\hat{V}_{imp}^{p} = \mathbb{E}_n\left[\eta_{imp}^{dr,p}\left(W; \hat{\gamma}^{ipt}, \hat{\beta}_{0,\Delta}^{wls,p}, \hat{\tau}_{imp}^{dr,p}\right)^2\right]$$

- for **repeated cross-section data**, again assume the working models

$$\pi(X,\gamma) = \Lambda\left(X'\gamma\right) \equiv \frac{\exp(X'\gamma)}{1+\exp(X'\gamma)} \qquad \mu_{d,t}^{rc}\left(X;\beta_{d,t}^{rc}\right) = \mu_{d,t}^{lin,rc}\left(X;\beta_{d,t}^{rc}\right) \equiv X'\beta_{d,t}^{rc}$$

then the two improved estimators are given as

$$\begin{aligned}
\hat{\tau}_{1,imp}^{dr,rc} =& \mathbb{E}_n\left[\left(\hat{w}_1^{rc}(D,T) - \hat{w}_0^{rc}\left(D,T,X;\hat{\gamma}^{ipt}\right)\right)\left(Y - \mu_{0,Y}^{lin,rc}\left(X;\hat{\beta}_{0,0}^{wls,rc},\hat{\beta}_{0,1}^{wls,rc}\right)\right)\right] \\
\hat{\tau}_{2,imp}^{dr,rc} =& \hat{\tau}_{1,imp}^{dr,rc} + \left(\mathbb{E}_n\left[\left(\frac{D}{\mathbb{E}_n[D]} - \hat{w}_{1,1}^{rc}(D,T)\right)\left(\mu_{1,1}^{rc}\left(X;\hat{\beta}_{1,1}^{ols,rc}\right) - \mu_{0,1}^{rc}\left(X;\hat{\beta}_{0,1}^{wls,rc}\right)\right)\right]\right) \\
& - \left(\mathbb{E}_n\left[\left(\frac{D}{\mathbb{E}_n[D]} - \hat{w}_{1,0}^{rc}(D,T)\right)\left(\mu_{1,0}^{rc}\left(X;\hat{\beta}_{1,0}^{ols,rc}\right) - \mu_{0,1}^{rc}\left(X;\hat{\beta}_{0,0}^{wls,rc}\right)\right)\right]\right)
\end{aligned}$$

where

$$\begin{aligned}
\hat{\gamma} &= \arg\max_{\gamma\in\Gamma} \mathbb{E}_n\left[DX'\gamma - (1-D)\exp(X'\gamma)\right] \\
\hat{\beta}_{0,t}^{wls,rc} &= \arg\min_{b\in\Theta} \mathbb{E}_n\left[\frac{\Lambda\left(X'\hat{\gamma}^{ipt}\right)}{1-\Lambda\left(X'\hat{\gamma}^{ipt}\right)}(Y - X'b)^2 \mid D=0, T=t\right] \\
\hat{\beta}_{1,t}^{ols,rc} &= \arg\min_{b\in\Theta} \mathbb{E}_n\left[(Y-X'b)^2 \mid D=1, T=t\right]
\end{aligned}$$

OLS is adopted to estimate $\beta_{1,t}^{*,rc}, t=0,1$ as there is no estimation effect.

Let

$$\tau_{imp}^{dr,rc} = \mathbb{E}\left[\left(w_1^{rc}(D,T) - w_0^{rc}\left(D,T,X;\gamma^{*,ipt}\right)\right)\left(Y - \mu_{0,Y}^{lin,rc}\left(T,X;\beta_{0,1}^{*,wls,rc},\beta_{0,0}^{*,wls,rc}\right)\right)\right]$$

and for $\beta_{imp}^{*,rc} = \left(\beta_{0,1}^{*,wls,rc}, \beta_{0,0}^{*,wls,rc}, \beta_{1,1}^{*,ols,rc}, \beta_{1,0}^{*,ols,rc}\right)$, define

$$\begin{aligned}
\eta_{1,imp}^{dr,rc}\left(W;\gamma^{*,ipt},\beta_{imp}^{*,rc}\right) &= \eta_1^{rc,1}\left(W;\beta_{0,1}^{*,wls,rc},\beta_{0,0}^{*,wls,rc}\right) - \eta_0^{rc,1}\left(W;\gamma^{*,ipt},\beta_{0,1}^{*,wls,rc},\beta_{0,0}^{*,wls,rc}\right) \\
\eta_{2,imp}^{dr,rc}\left(W;\gamma^{*,ipt},\beta_{imp}^{*,rc}\right) &= \eta_1^{rc,2}\left(W;\beta_{imp}^{*,rc}\right) - \eta_1^{rc,1}\eta_0^{rc,2}\left(W;\gamma^{*,ipt},\beta_{0,1}^{*,wls,rc},\beta_{0,0}^{*,wls,rc}\right)
\end{aligned}$$

Let $n = n_1 + n_0$, where n_1 and n_0 are the sample sizes of the post- and pre-treatment periods respectively. If $n_1/n \xrightarrow{p} \lambda \in (0,1)$ as $n_0, n_1 \to \infty$, then

$$\hat{\tau}^{dr,rc}_{j,imp} \xrightarrow{p} \tau, \quad \text{if } \textbf{either } \Lambda(X'\gamma^{*,ipt}) \overset{a.s.}{=} p(X) \textbf{ or } X'\beta^{*,wls,rc}_{0,1} - X'\beta^{*,wls,rc}_{0,0} \overset{a.s.}{=} m^{rc}_{0,\Delta}(X)$$

As for inference, $\hat{\tau}^{dr,p}_{imp}$ is $\sqrt{n}-$consistent and asymptotically normal

$$\sqrt{n}\left(\hat{\tau}^{dr,rc}_{j,imp} - \tau\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \eta^{dr,rc}_{j,imp}\left(W; \gamma^{*,ipt}, \beta^{*,rc}_{imp}\right) + o_p(1) \xrightarrow{d} \mathcal{N}\left(0, V^{rc}_{j,imp}\right)$$

and if **both** $\Lambda(X'\gamma^{*,ipt}) \overset{a.s.}{=} p(X)$ **and** $X'\beta^{*,wls,rc}_{d,t} \overset{a.s.}{=} m^{rc}_{d,t}(X)$, $\eta^{dr,rc}_{2,imp}\left(W; \gamma^{*,ipt}, \beta^{*,rc}_{imp}\right) \overset{a.s.}{=} \eta^{e,rc}(Y, D, T, X)$, $V^{rc}_{2,imp}$ equals to the semiparametrically efficiency bound, $V^{rc}_{1,imp}$ does **not**, where

$$V^{rc}_{j,imp} = \mathbb{E}_n\left[\eta^{dr,rc}_{j,imp}\left(W; \gamma^{*,ipt}, \beta^{*,rc}_{imp}\right)^2\right]$$

and the efficiency loss of using $\hat{\tau}^{dr,rc}_{1,imp}$ instaed of $\hat{\tau}^{dr,rc}_{2,imp}$ is

$$V^{rc}_{1,imp} - V^{rc}_{2,imp} = \mathbb{E}[D]^{-1} \cdot \text{Var}\left[\sqrt{\frac{1-\lambda}{\lambda}}\left(m^{rc}_{1,1}(X) - m^{rc}_{0,1}(X)\right) + \sqrt{\frac{\lambda}{1-\lambda}}\left(m^{rc}_{1,0}(X) - m^{rc}_{0,0}(X)\right) \mid D = 1\right] \geq 0$$

6.1.5.6 Unconfoundedness

Viewing the pre-treatment outcomes as covariates, then one can assume **unconfoundedness**:

$$D_i \perp (Y_{iT}(0), Y_{iT}(1)) \mid Y_{i1}, \cdots, Y_{iT-1}$$

under this assumption, one can apply the large literature of treatment effect estimation under unconfoundedness (Imbens, 2004) or modern approaches (Bang and Robins, 2005; Chernozhukov et al., 2017; Athey et al., 2018).

Imbens (2004) pointed out 3 arguments for the assumption of unconfoundedness

- statistical motivation: the unconfoundedness assumption is logically nature in program evaluation
- purpose: the unconfoundedness assumption asserts that all variables that need to be adjusted for are observed by the researcher
- even when agents choose their treatment optimally, two agents with the same values for observed characteristics may differ in their treatment choices without invalidating the unconfoundedness assumption if the difference in their choices is driven by differences in unobserved characteristics that are themselves unrelated to the outcomes of interest.

Imbens (2004) proposes that is is sufficient to assume a weaker version of unconfoundedness, **mean independence**

$$\mathbb{E}\left[Y_{iT}(0), Y_{iT}(1) \mid D_i, Y_{i1}, \cdots, Y_{iT-1}\right] = \mathbb{E}\left[Y_{iT}(0), Y_{iT}(1) \mid Y_{i1}, \cdots, Y_{iT-1}\right]$$

for population ATE.

Denote

$$\mu_d = \mathbb{E}_d(Y_{i1}, \cdots, Y_{iT-1}) = \mathbb{E}\left[Y \mid D_i = d, Y_{i1}, \cdots, Y_{iT-1}\right]$$

for $d = 0, 1$, Imbens (2004) reviews 5 groups of estimation for ATEs under unconfoundedness,

A. Regression for population/sample/conditional ATE we have the estimand

$$\hat{\tau}_{reg} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_1 - \hat{\mu}_0\right) = \frac{1}{N}\sum_{i=1}^{N} D_i \cdot \left(Y_i - \hat{\mu}_0\right) + (1 - D_i) \cdot \left(\hat{\mu}_1 - Y_i\right)$$

- **early estimators** for μ_d included parametric regression functions including least-square estimators with the regression function $\mu_d = \beta x + \tau \cdot d$, then one can use the regression

  $$Y_i = \alpha + \beta' \left(Y_{i1}, \cdots, Y_{iT-1}\right) + \tau W_i + \epsilon_i$$

  more generally, one can specify separate regressions for the two regimes $\mu_d = \beta_d x$, and estimate the two regressions separately.

  cons: the regression estimators may rely heavily on extrapolation, hence sensitive to changes in the specification of the models.
- **non-parametric estimators**
  - Hahn (1998) proposes to estimate first the 3 conditional expectations

    $$g_1(x) = \mathbb{E}\left[DY_t \mid Y_1, \cdots, Y_{T-1}\right] \quad g_0(x) = \mathbb{E}\left[(1-D)Y_t \mid Y_1, \cdots, Y_{T-1}\right] \quad e(x) = \mathbb{E}\left[D \mid Y_1, \cdots, Y_{T-1}\right]$$

    nonparametrically using series methods, then estimate[9]

    $$\hat{\mu}_1(x) = \frac{\hat{g}_1(x)}{\hat{e}(x)} \qquad\qquad \hat{\mu}_0(x) = \frac{\hat{g}_0(x)}{1 - \hat{e}(x)}$$

    and show that the estimators for population ATE and ATT achieve the semiparametric efficiency bounds. Alternatively, one can use this series approach to directly estimate μ_d.
  - Heckman et al. (1998a,b) propose (local-linear) kernel methods, one simple form is

    $$\hat{\mu}_d = \sum_{i:D_i=d} Y_{iT} \cdot K\left(\frac{X_i - x}{h}\right) \Big/ \sum_{i:D_i=d} K\left(\frac{X_i - x}{h}\right)$$

    with kernel $K(\cdot)$ and bandwidth h. In the local linear kernel regression, μ_d is estimated as the intercept β_0 in the minimization problem

    $$\min_{\beta_0,\beta_1} \sum_{i:D_i=d} \left[Y_i - \beta_0 - \beta_1'(X_i - x)\right]^2 \cdot K\left(\frac{X_i - x}{h}\right)$$

    for bias control, the order of the kernel should be at least as large as the dimension of the covariates: $\int_z z^r K(z)\mathrm{d}z = 0$, for $r \leq \dim(X)$.

for the population ATT, it is important to note that with the propensity score known, the average $\sum D_i Y_i / N_T$ is not efficient for the population expectation $\mathbb{E}\left[Y(1) \mid D = 1\right]$. The efficient estimator can instead be obtained by weighting all the estimated treatment effects $\hat{\mu}_1 - \hat{\mu}_0$ by the probability of receiving treatment:

$$\tilde{\tau}_{reg,T} = \frac{\sum_{i=1}^{N} e(X_i) \cdot \left[\hat{\mu}_1 - \hat{\mu}_0\right]}{\sum_{i=1}^{N} e(X_i)}$$

this allows one to exploit the control observations to adjust for imbalances in the sampling of the covariates.

[9] For simplicity, let $X_i \equiv (Y_{i1}, \cdots, Y_{iT-1})$.

B. Matching similar to nonparametric kernel regression methods, matching estimators also impute the missing potential outcomes, but using **only** the outcomes of **nearest neighbors** of the opposite treatment group[10]. Matching estimators often apply in settings where

- the interest is in the ATT
- there is a large reservoir of potential controls

the estimator is essentially the difference between 2 sample means, the variance is calculated using standard methods for differences in means or paired randomized experiments. The biggest challenge mostly comes from computation.

Abadie and Imbens (2002) propose that: again, for a sample $\{(Y_i, X_i, D_i)\}_{i=1}^N$, let $l_m(i)$ be the index that satisfies $D_l \neq D_i$, and

$$\sum_{j|D_j \neq D_i} \mathbf{1}\left\{\left\|X_j - X_i\right\| \leq \|X_l - X_i\|\right\} = m$$

then $l_m(i)$ is the index of the unit in the opposite treatment group that is the m^{th} closest to unit i based on norm $\|\cdot\|$ distance, and $l_1(i)$ is the nearest match. Let the set of indices for the first M matches for unit i be $\mathcal{L}_M(i) = \{l_1(i), \cdots, l_M(i)\}$, then define the imputed potential outcomes as

$$\hat{Y}_i(0) = \begin{cases} Y_i, & D_i = 0 \\ \frac{1}{M}\sum_{j\in\mathcal{L}_M(i)} Y_j, & D_i = 1 \end{cases} \qquad \hat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j\in\mathcal{L}_M(i)} Y_j, & D_i = 0 \\ Y_i, & D_i = 1 \end{cases}$$

the simple matching estimator is

$$\hat{\tau}_M^{sm} = \frac{1}{N}\sum_{i=1}^{N}\left[hat Y_i(1) - \hat{Y}_i(0)\right]$$

the bias of this estimator is $O\left(N^{-1/k}\right)$, where k is the dimension of the covariates. Imbens (2004) listed 3 caveats to Abadie and Imbens (2002)'s result:

1 only continuous covariates should be counted in k: matching with discrete covariates will be exact in large samples
2 if only matching the treated and the number of potential controls is much larger, the bias can be ignored asymptotically
3 the order of the bias may be high, but the actual bias may be small if the coefficients in the leading term are small[11]

and these matching estimators are generally not efficient.

One key aspect of matching is the choice of distance metrics, one can choose

- standard Euclidean metric: $d_E(x,z) = (x-z)'(x-z)$
- the diagonal matrix of the inverse of the covariate variances $d_{AI}(x,z) = (x-z)'\,\text{diag}\left(\Sigma_X^{-1}\right)(x-z)$
- Mahalanobis metric: $d_M(x,z) = (x-z)'\Sigma_X^{-1}(x-z)$, which reduces differences in covariates within matched pairs in all directions
- metrics depending on the correlation between covariates, treatment assignemnt and outcomes:
  - weighting absolute differences by the coefficients in the propensity score

$$d_{Z1}(x,z) = \sum_{k=1}^{K}|x_k - z_k|\cdot\left|\gamma_k\right|$$

[10]What makes matching estimators more attractive is that the researcher only has to choose the **number of matches**, instead of smoothing parameters.

[11]One such case is that biases for different units are at least partially offsetting.

where the propensity score has a logistic form $e(x) = \frac{\exp(x'\gamma)}{1+\exp(x'\gamma)}$

– weighting absolute differences by the coefficients in the regression function

$$d_{Z2}(x,z) = \sum_{k=1}^{K} |x_k - z_k| \left|\beta_k\right|$$

where the regression functions are linear $\mu_d(x) = \alpha_d + x'\beta$

Imbens (2004) comments that when the regression function is misspecified, matching with the particular metrics may lead to inconsistency.

C. Propensity scores There are 3 main ways to use propensity scores in estimation:

- **weighting**: the units by the reciprocal of the probability of receiving the treatment can undo the imbalance of the covariate distributions conditional on the treatment assignment. Formally,

$$\mathbb{E}\left[\frac{DY}{e(X)}\right] = \mathbb{E}\left[\frac{DY(1)}{e(X)}\right] = \underbrace{\mathbb{E}\left[\mathbb{E}\left[\frac{DY(1)}{e(X)} \mid X\right]\right] = \mathbb{E}\left[\frac{e(X)\cdot\mathbb{E}[Y(1)\mid X]}{e(X)}\right]}_{\text{unconfoundedness}} = \mathbb{E}[Y(1)]$$

and similarly $\mathbb{E}\left[\frac{(1-D)Y}{1-e(X)}\right] = \mathbb{E}[Y(0)]$, which imply

$$\tau^p = \mathbb{E}\left[\frac{D\cdot Y}{e(X)} - \frac{(1-D)\cdot Y}{1-e(X)}\right]$$

it can be directly estimated as

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_i Y_i}{e(X_i)} - \frac{(1-D_i)\cdot Y_i}{1-e(X_i)}\right]$$

here, the weights do not necessarily add to 1[12]. One can normalize the weights within subpopulations, which in the limits leads to the estimator proposed by Hirano et al. (2003)

$$\hat{\tau}_{weight} = \frac{\sum_{i=1}^{N}\frac{D_i\cdot Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^{N}\frac{D_i}{\hat{e}(X_i)}} - \frac{\sum_{i=1}^{N}\frac{(1-D_i)\cdot Y_i}{1-\hat{e}(X_i)}}{\sum_{i=1}^{N}\frac{1-D_i}{1-\hat{e}(X_i)}}$$

where they specify a sequence of functions of the covariates, such as power series $h_l(x), l = 1,\cdots,\infty$, and choose a number of terms $L(N)$ as a function of the sample size, then esitmate the $L-$dimensional vector γ_L in

$$\Pr(D=1\mid X=x) = \frac{\exp\left[(h_1(x),\cdots,h_L(x))\,\gamma_L\right]}{1+\exp\left[(h_1(x),\cdots,h_L(x))\,\gamma_L\right]}$$

by maximizing the associated likelihood function, and calculate the estimated propensity score as

$$\hat{e}(x) = \frac{\exp\left[(h_1(x),\cdots,h_L(x))\,\hat{\gamma}_L\right]}{1+\exp\left[(h_1(x),\cdots,h_L(x))\,\hat{\gamma}_L\right]}$$

a nonparametric estimator for $e(x)$ is efficient, ignoring the 2 regression functions[13].

[12]For the treated units, the weights added up to $\frac{1}{N}\sum_{i=1}^{N} W_i/e(X_i)$, which equals to 1 in expectation, but not so in any given sample.

[13]The finite-sample properties of the two approaches (nonparametrically estimating propensity scores versus regression functions) may be different, except for when there are only discrete covariates.

For the treatment effects of the treated, weight the contribution for unit i by the propensity score $e(x_i)$,

$$\hat{\tau}_{weight,tr} = \frac{\sum_{i=1}^{N} D_i \cdot Y_i \cdot \frac{e(X_i)}{\hat{e}(X_i)}}{\sum_{i=1}^{N} D_i \frac{e(X_i)}{\hat{e}(X_i)}} - \frac{\sum_{i=1}^{N}(1-D_i) \cdot Y_i \cdot \frac{e(X_i)}{(1-\hat{e}(X_i))}}{\sum_{i=1}^{N}(1-D_i)\frac{e(X_i)}{(1-\hat{e}(X_i))}} \quad \text{propensity scores known}$$

$$\hat{\tau}_{weight,tr} = \left[\frac{1}{N_1}\sum_{D_i=1} Y_i\right] - \left[\frac{\sum_{i:D_i=0} Y_i \cdot \frac{\hat{e}(X_i)}{(1-\hat{e}(X_i))}}{\sum_{i:D_i=0}\frac{\hat{e}(X_i)}{(1-\hat{e}(X_i))}}\right] \quad \text{propensity scores unknown}$$

CAUTIOUS: the problem of choosing the smoothing parameters is relevant here too[14], but here one want to use nonparametric regression methods even if the propensity scores are known.

- **blocking**: Rosenbaum and Rubin (1983) suggest using the (estimated) propsensity score divide the sample into M blocks of units of approximately equal probability of treatment, letting J_{im} be an indicator for unit i being in block m. One way of implementing this is by dividing the unit interval into M blocks with boundary values equal to m/M for $m = 1, \cdots, M-1$, s.t.

$$J_{im} = \mathbf{1}\left\{\frac{m-1}{M} < e(X_i) <\leq \frac{m}{M}\right\},\ m = 1, \cdots, M$$

within each block there are N_{dm} observations with treatment equal to d, $N_{dm} = \sum_i \mathbf{1}\{D_i = d, J_{im} = 1\}$. Within each block, estimate the average treatment effect as if random assignment held:

$$\hat{\tau}_m = \frac{1}{N_{1m}}\sum_{i=1}^{N} J_{im} W_i Y_i - \frac{1}{N_{0m}}\sum_{i=1}^{N} J_{im}(1-D_i) Y_i$$

then estimate the overall average treatment effect as

$$\hat{\tau}_{block} = \sum_{m=1}^{M} \hat{\tau}_m \cdot \frac{N_{1m} + N_{0m}}{N}$$

for the treatment effect of the treated, one can weight the within-block average treatment effects by the number of treated units:

$$\hat{\tau}_{T,block} = \sum_{m=1}^{M} \hat{\tau}_m \cdot \frac{N_{1m}}{N_T}$$

CAUTIOUS: The asymptotic properties of such estimators require establishing the relative relationship between the number of blocks and the sample size, so choosing the number of blocks becomes essential

- *starting point*: a single covariate, assuming normality, **5 blocks** removes $\geq 95\%$ of the bias
- *balance check*: covariates whould be balanced within blocks
- *unbalanced blocks*: if the distributions of the covariates among treated and controled are different, one can
  - split the blocks into a number of subblocks if the propensity score itself is unbalanced.
  - generalize the specification of the propensity score, if the score is balanced but covariates not
- *weighting with modified propensity score estimators*: discretize $\hat{e}(x)$ to

$$\tilde{e}(x) = \frac{1}{M}\sum_{m=1}^{M}\sum_{m=1}^{M} \mathbf{1}\left\{\frac{m}{M} \leq \hat{e}(x)\right\}$$

[14] Hirano, Imbens, and Ridder (2003)'s series estimators require choosing the **number of terms** in the series, a kernel-version alternative requires choosing a bandwidth.

then use $\tilde{e}(x)$ as the propensity score in the weighting estimator leads to an estimator for the ATE **identical** to that obtained by using the blocking estimator with $\hat{e}(x)$ as the propensity score and M blocks[15].

- **propensity scores as regressors**: estimate the conditional expectation of Y given D and $e(X)$, define

$$v_d(e) = \mathbb{E}[Y(d) \mid e(X) = e] \overset{\text{unconfoundedness}}{=} \mathbb{E}[Y \mid D = d, e(X) = e]$$

given an estimator $\hat{v}_w(e)$, one can estimate the ATE as

$$\hat{\tau}_{regprop} = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{v}_1\left(e(X_i)\right) - \hat{v}_0\left(e(X_i)\right)\right]$$

There are 2 cases in practice when considering propensity score approaches

- **propensity scores known**: all 3 methods are efficient, do not rely on high-dimensional nonparametric regressions, have attractive finite-sample properties
- **propensity scores unknown**: require high-dimensional nonparametric regression of the treatment indicator on the covariates. The relative merits of the 3 approaches will depend on whether the propensity score is more or less smooth than the regression functions, and whether additional information is available about either the propensity score or the regression functions.

D. Mixed Neither matching nor the propensity score methods directly address the correlation between the covariates and the outcome, incorporating the regression method may eliminate remaining bias and improve precision.

- **Weighting and Regression**: estimating

$$Y_i = \alpha + \tau \cdot D_i + \epsilon_i$$

with weights

$$\lambda_i = \sqrt{\frac{D_i}{e(X_i)} + \frac{1 - D_i}{1 - e(X_i)}}$$

the covariates are uncorrelated with the treatment indicator, making the weighted estimator consistent. To improve precision, add covariates

$$Y_i = \alpha + \beta' X_i + \tau \cdot D_i + \epsilon_i$$

which is doubly robust: consistent if either the regression model or the propensity score are specified correctly.

- **Blocking and Regression**: least square estimator in block m as

$$Y_i = \alpha_m + \tau_m \cdot D_i + \epsilon_i$$

using only units in block m. One can again add covariates and estimate $Y_i = \alpha_m + \beta_m' X_i + \tau_m \cdot D_i + \epsilon_i$.

- **Matching and Regression**: the bias of the simple matching estimator can dominate the variance if the dimension of the covariates is too large, regression can help in this situation.

Let $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ be the observed or imputed potential outcomes for unit i, the *estimated potential* outcomes equal the *observed* outcomes for some unit i for its match $l(i)$, the bias

$$\mathbb{E}\left[\hat{Y}_i(1) - \hat{Y}_i(0)\right] - \left[Y_i(1) - Y_i(0)\right]$$

[15]With sufficiently large M, the blocking estimator is sufficiently close to the oiriginal weighting estimator, sharing first-order asymptotic properties. Hence a large number of blocks does little harm, with regard to asymptotic properties.

arises from the fact that the covariates X_i and $X_{[}l(i)]$ for units i and $l(i)$ are not equal, although they are close because of the matching process. Focusing on the single-match case, define for unit i

$$\hat{X}_i(0) = \begin{cases} X_i & D_i = 0 \\ X_{l_1(i)} & D_1 = 1 \end{cases} \qquad \hat{X}_i(1) = \begin{cases} X_{l_1(i)} & D_i = 0 \\ X_i & D_1 = 1 \end{cases}$$

if the matching is exact, $\hat{X}_i(0) = \hat{X}_i(1)$ for each unit i. Suppose $D_1 = 1$, then $\hat{Y}_i(1) = Y_i(1)$, and $\hat{Y}_i(0)$ is an imputed value for $Y_i(0)$. This value is unbiased for $\mu_0\left(X_{l_1(i)}\right)$, but not necessarily for $\mu_0\left(X_i\right)$. Hence, $\hat{Y}_i(0)$ should be adjusted by an estimate of $\mu_0\left(X_i\right) - \mu_0\left(X_{l_1(i)}\right)$. Typically, these corrections are taken be to linear in the difference in the covariates for unit i and its match, of the form

$$\beta_0'\left[\hat{X}_i(1) - \hat{X}_i(0)\right] = \beta_0'\left(X_i - X_{l_1(i)}\right)$$

then Rubin (1973) proposed 3 modifications

1 using least squares, estimate

$$\hat{Y}_i\left(1\right) - \hat{Y}_(0) = \tau + \left[\hat{X}_i(1) - \hat{X}_i(0)\right]'\beta + \epsilon_i$$

2 estimate $\mu_0(x)$ directly by taking all control units and use lease squares to estimate

$$Y_i = \alpha_0 + \beta_0' X_i + \epsilon_i$$

if unit i is a control unit, or $Y_i = \alpha_1 + \beta_1' X_i + \epsilon_i$ for the treated units[16].

3 estimate the same regression function for the controls, but using **only those that are used as matches** for the treated units with weights corresponding to the number of times a control observation is used as a match.

CAUTIOUS: This approach can be less efficient due to dropping some control observations, but can likely avoid including outliners.

E. Bayesian Bayesian approaches have not been applied in estimating ATE under unconfoundedness. They can be useful for a number of reasons,

- **high-dimensionality**: Bayesian methods would allow researchers to include covariates with more or less informative prior distributions.
- **missing-at-random (MAR)**: multiple imputation methods often rely on a Bayesian approach for missing data.

Estimating variances

- for population ATE, the variance of efficient estamators is

$$V^P = \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} + \left(\mu_1(X) - \mu_0(X) - \tau\right)^2\right]$$

which can be estimated by

1 brute force: estimate all 5 components $\sigma_0^2(x)$, $\sigma_1^2(x)$, $\mu_0(x)$, $\mu_1(x)$ and $e(x)$ using kernel methods or series.
2 either estimate regression functions $\mu_0(x)$, $\mu_1(x)$ or the propensity score $e(x)$ using series/sieves. This can be interpreted as parametric estimators

[16] If the correction is done nonparametrically, the resulting matching estimator is consistent and asymptotically normal, with its bias dominated by the variance.

3 bootstrapping: given the asymptotic linearity of the estimators, bootstrapping will lead to valid standard errors and CIs at least for the regression and propensity score methods. For matching, it's more complicated since bootstrapping introduces discreteness in the distribution, which will lead to ties.

- for sample ATE, the appropriate (conservative) variance is

$$V^S = \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right]$$

which can be estimated by

1 estimating the conditional moments of the outcome distributions, with the accompanying inherent difficulties.

2 matching variance estimator[17]: the key is to obtain a close-to-unbiased estimator for $\sigma_w^2(x)$. Suppose units i and j have $X = x$, then an unbiased estimator for $\sigma_1^2(x)$ is

$$\hat{\sigma}_1^2(x) = \frac{1}{2}(Y_i - Y_j)^2$$

this matching doesn't need to be exact, one can use the closest match within the set of units with the same treatment indicator. Let $v_m(i)$ be the m-th closest unit to i with the same treatment indicator $W_{v_m(i)} = W_i$ and

$$\sum_{l|W_l=l, l\neq i} \mathbf{1}\left\{\|X_l - x\| \leq \left\|X_{v_m(i)} - x\right\|\right\} = m$$

which gives M units with the same treatment indicator and approximately the same values for the covariates. The sample variance of the outcome variable for these M units can then be used to estimate $\sigma_1^2(x)$, and similarly for the control variance function $\sigma_0^2(x)$.

3 estimate the variance of the ATE as

$$\hat{V}^S = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{\sigma}_1^2(X_i)}{\hat{e}(X_i)} + \frac{\hat{\sigma_0^2}(X_i)}{1-\hat{e}(X_i)}\right)$$

for matching estimators, even estimation of the propensity score can be avoided:

$$\hat{V}^E = \frac{1}{N}\sum_{i=1}^{N}\left(1 + \frac{K_M(i)}{M}\right)\hat{\sigma}_{W_i}^2(X_i)$$

where M is the number of matches and $K_M(i)$ is the number of times unit i is used as a match.

Assessing the assumptions There assumption of unconfoundedness is not directly testable, since the distribution of $Y(0)$ for those who received the treatment and of $Y(1)$ for those who received the control are never in the data. But there are still 2 groups of ways to indirectly test the assumption:

- estimating the casual effect of a treatment that is known **not to have** an effect: postulate a three-valued indicator $T_i \in \{-1, 0, 1\}$ for the groups of ineligibles, eligible non-participants and participants, and treatment indicator $W_i = \mathbf{1}\{T_i = 1\}$. The extended unconfoundedness assumption is

$$Y_i(0), Y_i(1) \perp T_i \mid X_i$$

and a testable implication is

$$Y_i \perp \mathbf{1}\{T_i = 0\} \mid X_i, T_i \leq 0$$

[17]The idea is that one need not actually estimate this variance consistently at all values of the covariates. One needs only the average of this variance over the distribution, weighted by the inverse of either $e(x)$ ot its complement $1 - e(x)$.

- estimating the casual effect of the treatment on a variable known to be **unaffected** by it, typically a pre-determined variable, which could eitehr me time-invariant, or more interestingly, a lagged outcome. One can test

$$Y_{i,-1} \perp W_i \mid Y_{i,-2}, \cdots, Y_{i,-T}, Z_i$$

which combines 2 assumptions[18]

$$Y_i(1), Y_i(0) \perp W_i \, Y_{i,-1}, \cdots, Y_{i,-(T-1)}, Z_i \qquad \text{unconfoundedness given only } T-1 \text{ lags}$$

$$f_{Y_{i,s(0)|Y_{i,s-1}(0)},\cdots,Y_{i,s-(T-1)}(0),Z_i,W_i}\left(y_s \mid y_{s-1}, \cdots, y_{s-(T-1)}, z, w\right) \qquad \text{stationarity and exchangeability}$$

next, the important question is how to **choose the covariates**, some concerns are

- some variables should not be adjusted for: the unconfoundedness assumption may apply with one set of covariates, but **not** with an expanded set. A covariate measured before the treatment was chosen should be safe to include in theory, but in practice, the covariates are often recorded as the same time as the outcomes
- the expected mean squared error may be reduced by ignoring those covariates that have only weak correlation with the treatment indicator and the outcomes: including a covariate in teh adjustment procedure will not lower the asymptotically precision of ATE but could reduce precision **in finite samples** if the covariates are not or only weakly correlated with outcomes and treatment indicators.

the second key assumption is the **overlapping**: the propensity score, i.e., the probability of receiving the active treatment, should be strictly between 0 and 1. In practice, there are 2 main issues

1 **detect** the lack of overlapping: there are several methods popularly used
   - plot distributions of covariates by treatment groups: can be very difficult in **high-dimensional** cases
   - non-parametrically estimate the distribution of the propensity score: one may wish to **undersmooth** the estimation by choosing a bandwidth smaller than optimal or including higher-order terms
   - inspect the **worst** matches: for each component k of the covariate vector, check

$$\max_i \left| x_{i,k} - x_{l_1(i),k} \right|$$

   the maximum over all observations of the matching discrepancy. If it's large relative to the **sample standard deviation** of the k-th component of the coviarates, there would be a lack of overlapping.

2 **address** the lack of overlapping: given a lack of overlap, one can
   - conclude the ATE **cannot** be estimated with sufficient precision
   - focus on an average treatment effect that is estimable with greater accuracy, by *dropping observations with a **cutoff** of propensity scores*: one principle is to evaluate the weight of each unit as

$$\frac{1}{N \cdot [1 - e(X_i)]}, \text{ for treated units} \qquad \frac{1}{N \cdot e(X_i)}, \text{ for control units}$$

   and make sure the weights of all units are **upper-bounded** by a reasonable number[19].

3 **comparing** the 3 approaches: in general, variance estimates **increase** when adding treated observations to a sufficiently-overlapping data set, approximately **unchanged** when adding control observations.
   - **regression**: in general, adding observations with outlier values of the regressors leads to more precise estimates. If the added observations are **treated** units, the precision of the estimated control regression at thes outlier values will be lower; if in **control** units, such precision will increase.

[18] The test depends on the link between the 2 assumptions and the original unconfoundedness assumption. With a sufficient number of lags, unconfoundedness given all lags but one appears plausible, conditional on unconfoundedness given all lags, so the relevance of the test depends largely on the plausiblity of the stationarity and exchangeability assumption.

[19] For example, set the cutoff as 0.05, s.t. no unit will have a weight of more than 5% in the average. In a sample with 1000 units, with such a cutoff, only units with a propensity score outside $[0.02, 0.98]$ will be ignored.

- **matching**: adding **control** observations with outlier covariate values will likely not change the results, such they won't be useed as matches; but adding **treated** will bias the estimates, while the standard error would largely unaffected.
- **propensity-score**: if the propensity score of a **control** unit is **close to 0**, adding it would not cause much trouble; but adding a **control** unit with a propensity score **close to 1** would increase variance of an ATE estimator

over all, **matching** and **propensity-score**, as well as **kernel-based** regression methods, are better in coping with limited overlap than (semi-)parametric regression models. In practice, one should *always* inspect histograms of the estimated propensity scores in both groups to assess whether limited overlap is an issue.

Chernozhukov et al. (2017) propose to a machine-learning based approach. They consider the case where treatment effects are fully heterogeneous, and the treatment variable $D \in \{0,1\}$, and model random vector (Y, D, Z) as

$$\begin{aligned} Y &= g_0(D,Z) + \zeta & \mathbb{E}[\zeta \mid Z, D] &= 0 & \text{outcome variable} \\ D &= m_0(Z) + v & \mathbb{E}[v \mid Z] &= 0 & \text{treatment assignment} \end{aligned}$$

which allows for general heterogeneity in treatment effect. The confounding factors Z affect the treatment variable D via propensity score $m_0(Z) \coloneqq \mathbb{E}[D \mid Z]$, and the outcome variable via function $g_0(D,Z)$, both of these functions will be estimated via ML methods.

Then the ATE or ATT are

$$\begin{aligned} \theta_0 &= \mathbb{E}\left[g_0(1,Z), g_0(0,Z)\right] & \text{ATE} \\ \theta_0 &= \mathbb{E}\left[g_0(1-Z) - g_0(0,Z) \mid D=1\right] & \text{ATT} \end{aligned}$$

conisder a score $\psi(W;\theta,\eta)$ that satisfies

$$\begin{aligned} \mathbb{E}\psi\left(W;\theta_0,\eta_0\right) &= 0 & \text{identification condition} \\ \partial_n \mathbb{E}\psi\left(W;\theta_0,\eta\right)\big|_{\eta=\eta_0} &= 0 & \text{Neyman orthogonality condition} \end{aligned}$$

Chernozhukov et al. (2017) suggest employing

- for **ATE**

$$\begin{aligned} \psi\left(W;\theta,\eta\right) &= \frac{D\left(Y-g(0,Z)\right)}{m} - \frac{m(Z)(1-D)\left(Y-g(0,Z)\right)}{(1-m(Z))\,m} - \theta\frac{D}{m} \\ \eta(Z) &\coloneqq \left(g(0,Z), g(1,Z), m(Z)\right) \\ \eta_0(Z) &\coloneqq \left(g_0(0,Z), g_0(1,Z), m_0(Z)\right) \end{aligned}$$

- for **ATT**

$$\begin{aligned} \psi\left(W;\theta,\eta\right) &= \frac{D\left(Y-g(0,Z)\right)}{m} - \frac{m(Z)(1-D)\left(Y-g(0,Z)\right)}{(1-m(Z))\,m} - \theta\frac{D}{m} \\ \eta(Z) &\coloneqq \left(g(0,Z), g(1,Z), m(Z), m\right) \\ \eta_0(Z) &\coloneqq \left(g_0(0,Z), g_0(1,Z), m_0(Z), \mathbb{E}[D]\right) \end{aligned}$$

where $\eta(Z)$ is the nuisance parameter with true value of $\eta_0(Z)$ consisting of $P-$square integrable functions, mapping the support of Z to $\mathbb{R}\times\mathbb{R}\times(\epsilon,1-\epsilon)$ for ATE and to $\mathbb{R}\times\mathbb{R}\times(\epsilon,1-\epsilon)\times(\epsilon,1-\epsilon)$ for ATT[20].

[20] All semiparametrically efficient scores satisfy Neyman orthogonality, but not vice versa. In some problems, one may consider inefficient orthogonal scores for robustness.

Chernozhukov et al. (2017) propose to use cross-fitting to avoid overfitting and the Neyman orthogonality to reduce regularization and modeling biases of ML estimators $\hat{\eta}_0$, hence, **double debiasing**

Algorithm 6.1.9: Estimation with Orthogonal Scores by $K-$fold Cross-Fitting

S1 Let K be a fixed integer. Form a $K-$fold random partition of $\{1, \cdots, N\}$ by dividing it into equal parts $(I_k)_{k=1}^K$, each of size $n := N/K$. For each I_k, let I_k^C denote all indices that are **not** in I_k

S2 Construct K estimators $\check{\theta}_0\left(I_k, I_k^C\right), k = 1, \cdots, K$ that employ the machine learning estimators

$$\hat{\eta}_0\left(I_k^C\right) = \left(\hat{g}_0\left(0, Z; I_k^C\right), \hat{g}_0\left(1, Z; I_k^C\right), \hat{m}_0\left(Z; I_k^C\right), \frac{1}{N-n}\sum_{i=\in I_k^C} D_i\right)'$$

of the nuisance parameters

$$\eta_0(Z) = \left(g_0(0, Z), g_0(1, Z), m_0(Z), \mathbb{E}[D]\right)'$$

and where each estimator $\check{\theta}_0\left(I_k, I_k^C\right)$ is defined as the root θ of

$$\frac{1}{n}\sum_{i\in I_k}\psi\left(W; \theta, \hat{\eta}_0\left(I_k^C\right)\right) = 0$$

for the score ψ for ATE and ATT respectively.

- **S3** Average the K estimators to obtain the final estimator

$$\tilde{\theta}_0 = \frac{1}{K}\sum_{k=1}^K \check{\theta}_0\left(I_k, I_k^C\right)$$

and an approximate standard error for this estimator is $\hat{\sigma}/\sqrt{N}$, where $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^N \hat{\psi}_i^2$, $\hat{\psi} := \psi\left(W_i; \tilde{\theta}_0, \hat{\eta}_0\left(I_{k(i)}^C\right)\right)$, $k(i) := \{k \in \{1, \cdots, K\} : i \in I_k\}$ an approximate $(1-\alpha)\times 100\%$ Confidence interval is

$$\mathrm{CI}_n := \left[\tilde{\theta}_0 \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\hat{\sigma}}{\sqrt{N}}\right]$$

The **key assumptions** on the rate of estimating are

- $\forall d \in \{0,1\}$, $\|\zeta\|_{P,2} \geq c$, $\|v\|_{P,2} \geq c$, and

$$\left\|g(d, Z)\right\|_{P,q} \leq C \qquad \|Y\|_{P,q} \leq C \qquad P\left(\epsilon \leq m_0(Z) \leq 1-\epsilon\right) = 1 \qquad P\left(\mathbb{E}_P\left[\zeta^2 \mid Z\right] \leq C\right) = 1$$

- the ML estimators based on a random subset I_k^C of $\{1, \cdots, N\}$ of size $N-n$, $\forall N \geq 2K, d \in \{0,1\}$

$$\left\|\hat{g}_0\left(d, Z; I_k^C\right) - g_0(d, Z)\right\|_{P,2} \cdot \left\|\hat{m}_0\left(Z; I_k^C\right) - m_0(Z)\right\|_{P,2} \leq \frac{\delta_n}{\sqrt{n}}$$

$$\left\|\hat{g}_0\left(d, Z; I_k^C\right) - g_0(d, Z)\right\|_{P,2} + \left\|\hat{m}_0\left(Z; I_k^C\right) - m_0(Z)\right\|_{P,2} \leq \delta_n$$

and $P\left(\epsilon \leq \hat{m}_0\left(Z; I_k^C\right) \leq 1-\epsilon\right) = 1$ with P_P-probability no less than $1-\Delta_n$[a].

[a] The conditions are fairly sharp, in the sense that for a sparse regression function g_0 and a propsensity function m_0 with sparsity indices $s^g \ll n, s^m \ll n$, and estimators $\hat{g}_0$ and $\hat{m}_0$ have sparsity indices of order s^g and s^m, converging to g_0 and m_0 at the rates $\sqrt{s^g/n}$ and $\sqrt{s^m/n}$. The rate conditions in the assumption require

$$\sqrt{s^g/n}\cdot\sqrt{s^m/n} \Leftrightarrow s^g s^m \ll \sqrt{n}$$

which is much weaker than the without-sample-splitting condition $(s^g)^2 + (s^m)^2 \ll n$

the estimator from Algorithm 6.1.9 achieves asymptotic normality: $\sigma^{-1}\sqrt{N}\left(\tilde{\theta}_0 - \theta_0\right) \to \mathcal{N}(0,1)$, where $\sigma^2 = \mathbb{E}_P\left[\psi^2\left(W; \theta_0, \eta_0(Z)\right)\right]$, the results hold with $\hat{\sigma}^2$. The confidence regions based upon $\tilde{\theta}_0$ have uniform asymptotic validity

$$\sup_{P \in \mathcal{P}} \left| P\left(\theta_0 \in CI_n\right) - (1-\alpha) \right| \to 0$$

.

Uncertainty due to sample-splitting: although sample partitions have no impact on estimation results asymptotically, but may be important in finite samples. Chernozhukov et al. (2017) proposes to repeat the procedure S times and repartition each time, get a set of estimates $\tilde{\theta}_0^s$. Then, consider 2 quantities:

- sample **average** of the S estimates: $\tilde{\theta}_0^{\text{mean}}$
- sample **median** of the S estimates: $\tilde{\theta}_0^{\text{median}}$

$\tilde{\theta}_0^{\text{median}}$ is less affected by extreme estimates, hence more robust. But asymptotically, the specific random partition is irrelevant, hence $\tilde{\theta}_0^{\text{mean}}$ and $\tilde{\theta}_0^{\text{median}}$ should be close. Chernozhukov et al. (2017) also introduce the repsective standard error measures:

$$\hat{\sigma}^{\text{mean}} = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\hat{\sigma}_s^2 + \underbrace{\left(\tilde{\theta}_0^s - \frac{1}{S}\sum_{j=1}^{S}\tilde{\theta}_0^j\right)^2}_{\text{variation of sample splitting}}\right)} \qquad \hat{\sigma}^{\text{median}} = \text{median}\left\{\sqrt{\hat{\sigma}_i^2 + \underbrace{\left(\hat{\theta}_i - \hat{\theta}^{\text{median}}\right)^2}_{\text{variation of sample splitting}}}\right\}_{i=1}^{S}$$

In practice, one might want to consider the following

- for **extreme propensity score**, set cutoffs close to 0 and 1
- for **nuisance function estimation**, some typical methods are tree-based methods (Random Forest, Regression Tree, Boosting), or l_1-penalization method (Lasso), or neural network. One can also use the weighted averages of these methods s.t. the average gives the lowest mean squared out-of-sample prediction errors.

Athey et al. (2018) proposed a similar framework, focusing on high dimensional linear models. They showed that in linear models, it is enough to correct for linear biases in two steps:

S1 Fit a regularized linear model for the outcome given the features separately in the 2 treatment groups

S2 reweight the first-stage residuals by using weights that approximately balance all the features between the treatment and control groups

Athey et al. (2018) showed that it is often possible to achieve approximate balance under reasonable assumptions and that approximate balance suffices for eliminating bias due to regularization, combined with a lasso regression adjustment.

Consider the conditional ATE for the treated sample

$$\tau = \frac{1}{n_t}\sum_{i:D_i=1}\mathbb{E}\left[Y_i(0) - Y_i(1) \mid X_i\right]$$

In addition to unconfoundedness, they also assume **linearity**:

$$\mu_c(x) = \mathbb{E}\left[Y_i(0) \mid X = x\right] = x\beta_c \qquad \mu_t(x) = \mathbb{E}\left[Y_i(1) \mid X = x\right] = x\beta_t$$

for all $x \in \mathbb{R}^p$. This assumption is strong, but generally used in high-dimension cases. Given linearity, we have

$$\tau = \mu_t - \mu_c \qquad \mu_t = \overline{X}_t \beta_t \qquad \mu_c = \overline{X}_t \beta_c \qquad \overline{X}_t = \frac{1}{n_t} \sum_{i=1}^{n} \mathbf{1}\{D_i = 1\} X_i$$

the main focus of this paper is to estimate μ_c when p is large, combining two approaches:

- **weighted** estimation: weighting control to mimic treatment
- **regression** adjustment: estimate β_c using control observations and get $\hat{\mu}_c = \overline{X}_t \hat{\beta}_c$

both of them perform poorly in high-dimension setting alone. Athey et al. (2018) propose a meta-algorithm that estimate

$$\hat{\mu}_c = \overline{X}_t \hat{\beta}_c + \sum_{i:W_i=0} \gamma_i \left(Y_i^{obs} - X_i \hat{\beta}_c \right)$$

where $\overline{X}_t \hat{\beta}_c$ captures the strong signals, and γ_i rebalance the residuals and effectively extract left-over signals. It satisfies

$$\left| \hat{\mu}_c - \mu_c \right| \leq \underbrace{\left\| \overline{X}_t - X_c^T \gamma \right\|_\infty \left\| \hat{\beta}_c - \beta_c \right\|_1}_{\text{dimensionality bias}} + \underbrace{\left| \Sigma_{i:W_i=0} \gamma_i \epsilon_i \right|}_{\text{variance}}$$

the dimensionality bias should be expected to scale as

$$\left\| \overline{X}_t - X_c^T \gamma \right\|_\infty = O\left(\sqrt{\log(p)/n} \right) \qquad \left\| \hat{\beta}_c - \beta_c \right\|_1 = O\left(k \cdot \sqrt{\log(p)/n} \right)$$

then in a sufficiently sparse setting, i.e., k is sufficiently small, the bias could be **variance dominated**. The estimation procedure is

S1 compute the **positive** approximately balancing weights γ to make the mean of the reweighted control sample $X_c^T \gamma$ match the treated sample mean $\overline{X}_t$ as closely as possible,

$$\gamma = \arg\min_{\tilde{\gamma}} \left\{ (1-\zeta) \left\| \tilde{\gamma} \right\|_2^2 + \zeta \left\| \overline{X}_t - X_c^T \tilde{\gamma} \right\|_\infty^2 \text{ s.t. } \sum_{i:W_i=0} \tilde{\gamma}_i = 1, 0 \leq \tilde{\gamma}_i \leq n_c^{-2/3} \right\}$$

γ is constrained to be positive to prevent aggresive extrapolating.

S2 fit β_c in the linear model by using a lasso or an elastic net to achieve sufficiently good risk bounds under L_1-risk

$$\hat{\beta}_c = \arg\min_{\beta} \left[\sum_{i:W_i=0} \left(Y_i^{obs} - X_i \beta \right)^2 + \lambda \left\{ (1-\alpha) \left\| \beta \right\|_2^2 + \alpha \left\| \beta \right\|_1 \right\} \right]$$

S3 estimate the ATE as

$$\hat{\tau} = \overline{Y}_t - \left\{ \overline{X}_t \hat{\beta}_c + \sum_{i:W_i=0} \gamma_i \left(Y_i^{obs} - X_i \hat{\beta}_c \right) \right\}$$

an analogous estimator for the ATE $\mathbb{E}[Y(1) - Y(0)]$ can also be constructed as

$$\hat{\tau}_{ATE} = \overline{X} \left(\hat{\beta}_t - \hat{\beta}_c \right) + \sum_{i:W_i=1} \gamma_{t,i} \left(Y_i^{obs} - X_i \hat{\beta}_t \right) - \sum_{i:W_i=0} \gamma_{c,i} \left(Y_i^{obs} - X_i \hat{\beta}_c \right)$$

where

$$\gamma_t = \arg\min_{\tilde{\gamma}} \left\{ (1-\zeta) \left\| \tilde{\gamma} \right\|_2^2 + \zeta \left\| \overline{X} - X_t^T \tilde{\gamma} \right\|_\infty^2 \text{ s.t. } \sum_{i:W_i=0} \tilde{\gamma}_i = 1, 0 \leq \tilde{\gamma}_i \leq n_c^{-2/3} \right\}$$

and γ_c is constructed similarly.

There are 3 key features of this algorithm:

(a) direct covariance adjustment based on the outcome data with regularization to deal with the **high dimensionality**
(b) **weighting** using the relationship between the treatment and the features
(c) the weights are based on **direct measures of imbalance** rather than on propensity score estimates

In comparison with doubly robust estimators,

- doubly robust estimators estimate the outcome model and the propensity score, this method instead relies on linearity assumption twice
- doubly robust estimators are preferable given sufficient sparsity and well-specified propensity models
- this method is not estimating propensity score, but just balancing, hence substantially easier

Unconfoundedness versus DiD/TWFE The unconfoundedness assumption and TWFE model validate different non-nested comparisons:

- **TWFE** model assumes: the treated differs from the control in unobserved characteristics that are potentially **correlated** with a persistent component of the outcomes (the fixed effects α_i)
- **Unconfoundedness** assumes: the selection is based **solely on past** rather than future outcomes

The literature in general does not provide a lot of guidance on the choice between the two strategies[21]. In two cases, unconfoundedness and TWFE lead to similar results:

- the averages of previous period outcomes are similar for the control and the treatment group
- the avearge in the control group does not change much over time

when the control group changes over time, and the control group and treatment group differ in the initial period, then TWFE and unconfoundedness give different results. The differences can be bounded under additional assumptions. Consider the case of 2 periods, where everyone in the first period with treatment equal to 0, and a positive true treatment effect β.

- If treatment is correlated with an unobserved individual fixed effects a_i, and outcomes are

  $$Y_{it} = a_i + \beta D_{it} + \epsilon_{it} \qquad Y_{it-1} = a_i + \epsilon_{it-1}$$

  where ϵ_{it} is serially uncorrelated, and uncorrelated with a_i and D_{it}. If one controls for Y_{it-1} but ignore fixed effects, that is, estimate Y_{it} on the residual from a regression of D_{it} on Y_{it-1}, the resulting estimator is $\frac{\text{Cov}(Y_{it}, D_{it} - \gamma Y_{it-1})}{\text{Var}(D_{it} - \gamma Y_{it-1})}$. Substituting $a_i \equiv Y_{it-1} - \epsilon_{it-1}$, get the real

  $$Y_{it} = Y_{it-1} + \beta D_{it} + \epsilon_{it} - \epsilon_{it-1}$$

  then the estimator controlling unconfoundedness only is

  $$\frac{\text{Cov}\left(Y_{it}, D_{it} - \gamma Y_{it-1}\right)}{\text{Var}\left(D_{it} - \gamma Y_{it-1}\right)} = \beta - \frac{\text{Cov}\left(\epsilon_{it-1}, D_{it} - \gamma Y_{it-1}\right)}{\text{Var}\left(D_{it} - \gamma Y_{it-1}\right)} = \beta + \gamma \cdot \frac{\sigma_\epsilon^2}{\text{Var}\left(D_{it} - \gamma Y_{it-1}\right)}$$

[21] Xu (2023) touches on this topic by comparing **strict exogeneity** (TWFE) and **sequential ignorability** (unconfoundedness).

if Y_{it-1} is larger in the control group, meaning that $\gamma < 0$, just assumming unconfoundedness and adjusting for the lagged outcome will overestimate the true effects.

- If the correct specification is instead

$$Y_{it} = \alpha + \theta Y_{it-1} + \beta D_{it} + \epsilon_{it}$$

where ϵ_{it} is serially uncorrelated. Just run a TWFE (in this case, first-difference) model, get an estimator $\frac{\text{Cov}(Y_{it}-Y_{it-1},D_{it})}{\text{Var}(D_{it})}$, plug in $Y_{it} - Y_{it-1} = \alpha + (\theta - 1) Y_{it-1} + \beta D_{it} + \epsilon_{it}$, get

$$\frac{\text{Cov}\left(Y_{it} - Y_{it-1}, D_{it}\right)}{\text{Var}\left(D_{it}\right)} = \beta + (\theta - 1) \cdot \frac{\text{Cov}\left(Y_{it-1}, D_{it}\right)}{\text{Var}\left(D_{it}\right)}$$

in general $\theta \in (0, 1)$ for stationary, if Y_{it-1} is larger in control group, TWFE will overestimate the true effects.

In the more general setting, to estimate $\mu_0 = \mathbb{E}\left[Y_{i,t+1}(0) \mid G_i = 1\right]$, we are comparing

$$\tilde{\mu}_{0,TWFE} = \mathbb{E}\left(Y_{it} \mid G_i = 1\right) + \mathbb{E}\left(Y_{i,t+1} \mid G_i = 0\right) - \mathbb{E}\left(Y_i t \mid G_i = 0\right) \quad \text{TWFE}$$

$$\tilde{\mu}_{0,uncf} = \mathbb{E}\left[\mathbb{E}\left(Y_{t+1} \mid G = 0, Y_t\right) \mid G = 1\right] = \int \mathbb{E}\left(Y_{t+1} \mid G = 0, Y_t = y\right) F_{Y_t}\left(\mathrm{d}y \mid G = 1\right) \quad \text{unconfoundedness}$$

the difference between the two is then

$$\tilde{\mu}_{0,uncf} - \tilde{\mu}_{0,TWFE} = \int \Delta(y) F_{Y_t}\left(\mathrm{d}y \mid = 1\right) - \int \Delta(y) F_{Y_t}\left(\mathrm{d}y \mid G = 0\right)$$

where $\Delta(y) = \mathbb{E}\left(Y_{t+1} \mid G = 0, Y_t = y\right) - y = \mathbb{E}\left(Y_{t+1} - Y_t \mid G = 0, Y_t = y\right)$ equals the expectation of change in the outcome conditioning on the lagged outcome in the **control** group. Ding and Li (2019) establishes that the relative magnitude of $\tilde{\mu}_{0,uncf}$ and $\tilde{\mu}_{0,TWFE}$ depends

(a) $\mathbb{E}\left(Y_{t+1} - Y_t\right)$ conditional on Y_t in the **control** group: the dependence of the outcome on the lagged outcome

(b) difference between the distribution of Y_t in the **treated** versus **control** group: the dependence of the treatment assignment on the lagged outcome

Assuming stationarity,

$$\frac{\partial \mathbb{E}\left(Y_{t+1} \mid G = 0, Y_t = y\right)}{\partial y} < 1, \forall y$$

we have

- if $Y_t \perp G$, or equivalently, $F_{Y_t}\left(y \mid G = 1\right) = F_{Y_t}\left(y \mid G = 0\right)$
- if $F_{Y_t}(y \mid G = 1) \geq F_{Y_t}(y \mid G = 0), \forall y$, $\tilde{\mu}_{0,TWFE} \leq \tilde{\mu}_{0,uncf}$, thus $\tilde{\tau}_{TWFE} \geq \tilde{\tau}_{uncf}$
- if $F_{Y_t}(y \mid G = 1) \leq F_{Y_t}(y \mid G = 0), \forall y$, $\tilde{\mu}_{0,TWFE} \geq \tilde{\mu}_{0,uncf}$, thus $\tilde{\tau}_{TWFE} \leq \tilde{\tau}_{uncf}$

and both estimation could be biased for the true τ_{ATT}[22].

For the semiparametric inverse probability weighting estimator proposed by Abadie (2005),

$$\tilde{\mu}'_{0,TWFE} = \mathbb{E}\left[GY_t + \frac{e(1-G)\left(Y_{t+1} - Y_t\right)}{1-e}\right] / \Pr\left(G = 1\right)$$

[22]Stationary can be tested by estimating the derivative of the conditional mean function $\mathbb{E}\left(Y_{t+1} \mid G = 0, Y_t = y\right)$; $F_{Y_t}\left(y \mid G = 1\right)$ and $F_{Y_t}\left(y \mid G = 0\right)$ can be visually compared via the empirical CDF of the outcomes in the treatment and control groups.

and the semiparametric estimator under unconfoundedness

$$\tilde{\mu}'_{0,uncf} = \mathbb{E}\left[\frac{e(Y_t)}{1-e(Y_t)}(1-G)Y_{t+1}\right] / \Pr(G=1)$$

the comparison results still hold.

Imai et al. (2023) propose a TWFE/DID mixed method by conditioning on lagged outcomes other than the most recent one, which is differenced out by TWFE. For each treated observation,

1 find a set of control observations that have the idential treatment history up to the prespecified number of periods
2 refine the *matched set* by adjusting for observed confounding via standard matching and weighting techniques s.t. the treated and matched control observations have similar covariate values
3 apply a DiD estimator to account for an underlying time trend

Imai et al. (2023) defined 2 parameters

- the number of *leads*, $F \geq 0$: how many periods **after** the treatment being administered, selected by researchers
- the number of *lags*, $L \geq 0$: how many periods **before** the treatment being administered, selected based on identification assumption

References

Alberto Abadie. Semiparametric difference-in-differences estimators. *The review of economic studies*, 72(1): 1–19, 2005.

Alberto Abadie and Guido Imbens. Simple and bias-corrected matching estimators for average treatment effects, 2002.

Dmitry Arkhangelsky and Guido Imbens. Causal models for longitudinal and panel data: A survey. Technical report, National Bureau of Economic Research, 2023.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):597–623, 2018.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107 (5):261–265, 2017.

Peng Ding and Fan Li. A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis*, 27(4):605–615, 2019.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

James Heckman, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. Characterizing selection bias using experimental data. *Econometrica*, pages 1017–1098, 1998a.

James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294, 1998b.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Kosuke Imai, In Song Kim, and Erik H Wang. Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 67(3):587–605, 2023.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, pages 185–203, 1973.

Pedro HC Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of econometrics*, 219(1):101–122, 2020.

Yiqing Xu. Causal inference with time-series cross-sectional data: a reflection. *Available at SSRN 3979613*, 2023.