Прикладная эконометрика, 2024, т. 73, с. 119–142.
Applied Econometrics, 2024, v. 73, pp. 119–142.
DOI: 10.22394/1993-7601-2024-73-119-142

Е.С. Котырло[1]

Простой и сложный метод разности разностей

В статье обсуждается применение метода разности разностей (difference-in-differences, DD) для оценки эффекта регулирующего воздействия, который, благодаря своей популярности, получил развитие со случая двух периодов и двух сравниваемых групп на случай нескольких периодов, нескольких групп, воздействия, которое прекращается и возобновляется, вероятного воздействия и квантильного эффекта воздействия. В статье кратко описаны случаи, когда получение состоятельной оценки эффекта воздействия возможно в классической модели для панельных данных с фиксированными временными и индивидуальными эффектами (TWFE), а когда оправдано применение альтернативных методов. Статья кратко представляет методы, развивающие классический подход к оценке DD с использованием TWFE модели, а также непараметрические и полупараметрические подходы. Известное условие параллельных трендов (PTA) в некоторых методах может быть заменено на условное PTA или случайное по времени воздействие. В статье даются ссылки на реализации этих методов в Stata и R. Имитационное моделирование демонстрирует, что декларируемые разработчиками новых методов свойства не всегда выполняются.

Ключевые слова: метод разности разностей; условие параллельных трендов; непоследовательное воздействие; вероятное воздействие; квантильный эффект воздействия.

JEL classification: C18; C23; C87.

1. Введение

Метод разности разностей (difference-in-differences, double difference, DID, далее DD), предложенный в работах (Ashenfelter, 1978; Ashenfelter, Card, 1985), стал одним из самых популярных эконометрических подходов к оценке реформ и регулирующих воздействий. Со времени последней обзорной статьи об этом методе, опубликованной на русском языке (Вулдридж, 2009), метод получил развитие в разных направлениях. Разнообразие альтернатив создает трудности выбора не только для практиков, но и для знатоков эконометрики. Непросто понять, когда и какой вариант DD следует применять, и как разные подходы дополняют друг друга. На необходимость подвести итоги в развитии метода указывают даже названия недавних статей по этой теме: "…A synthesis of the recent econometrics literature" (Roth et al., 2023), "…differences-in-differences…: A survey" (De Chaisemartin, D'Haultfoeuille, 2022).

[1] **Котырло Елена Станиславовна** — Национальный исследовательский университет «Высшая школа экономики», Москва; ekotyrlo@hse.ru.

В настоящей статье приведены основные предположения, необходимые для применения DD, примеры их нарушения в случае нескольких периодов и групп воздействия, вероятного воздействия и квантильного эффекта воздействия. Описаны подходы, позволяющие расширить применение DD, ослабив эти предположения. Цель статьи — дать общие представления о том, когда можно и когда нельзя использовать классический метод оценки с фиксированными эффектами на панельных данных, а также указать недавнюю литературу (преимущественно на английском языке), где рассматриваются подходы к применению DD для оценки эффекта с учетом особенностей регулирующего воздействия и объектов, на которые оно направлено.

Статья структурирована следующим образом. Следующий раздел 2 знакомит читателя с базовой TWFE моделью 2×2 для измерения эффекта методом DD, концепцией среднего эффекта воздействия ATE (the average treatment effect) и среднего эффекта воздействия на группу воздействия ATT (the average treatment effect for the treated). В разделе 3 сформулирована общая модель для случая нескольких групп и периодов. Рассматривается состоятельность оценки, когда эффект не зависит от длительности воздействия (статический эффект) или меняется по времени и по группам воздействия (динамический эффект). Раздел 4 кратко представляет подходы к развитию модели TWFE на случай воздействия, неоднородного по времени и группам, предложенные Borusyak et al. (2021), Chaisemartin, D'Haultfoeuille (2020), Freyaldenhoven et al. (2021), Gardner (2021) и Sun, Abraham (2021). Раздел 5 кратко описывает методы, предложенные Callaway, Sant'Anna (2021), Chaisemartin, D'Haultfoeuille (2020), Roth, Sant'Anna (2021), позволяющие существенно ослабить предположения, необходимые для применения DD. В разделе 6 приводятся результаты имитационного моделирования для статического эффекта воздействия, реализующего подходы Borusyak et al. (2021), Callaway, Sant'Anna (2021), Freyaldenhoven et al. (2021), Gardner (2021), Sun и Abraham (2021), и Roth, Sant'Anna (2021). Последний раздел представляет основные выводы данной работы.

2. Оценка эффекта воздействия в модели DD 2×2

Модель 2×2. Базовая TWFE модель для измерения эффекта воздействия методом DD предполагает наличие двух периодов и двух групп, которые сравниваются между собой. Одна группа испытывает регулирующее воздействие (*Treated*), другая служит для сравнения (*Control*). Рассмотрим на примере, почему важно иметь группы, испытывающие и не испытывающие воздействие, и периоды до и после воздействия. Если сравнивать результирующие показатели в двух группах только в момент воздействия, то такое сравнение, очевидно, будет страдать от различия в наблюдаемых и ненаблюдаемых характеристиках сравниваемых групп. Рассмотрим это на примере.

Все не так давно пережили эпидемию коронавируса и связанные с ней ограничения на условия работы[2]. Все работники напрямую или косвенно испытывали воздействие коронавирусных ограничений, причем разнонаправленное. Для кого-то условия занятости

[2] Конечно, воздействие ограничений включает множество социально-экономических, психологических и прочих аспектов, но для простоты сосредоточимся только на одном.

ухудшились[3], а для кого-то улучшились[4]. В этом приводимый пример не вполне соответствует допущениям, необходимым для оценки эффекта (см. ниже), но в реальной экономике мало ситуаций, идеально подходящих под условия в теории. Поэтому предположим, что на рынке труда есть группы, которые испытывают воздействие ограничений в значительно большей степени, чем другие, и группы, сохранившие стабильную занятость и заработок в период COVID-19. К первым отнесем индивидов, чей трудовой контракт не гарантирует долгосрочной занятости, например, занятые без оформления трудовых отношений. Скорее всего, в этой группе много неквалифицированных работников, студентов, подрабатывающих во время учебы на условиях неполной занятости, и т.д. Будем считать их группой воздействия. А кто испытал наименьшее воздействие? Это работники, защищенные трудовым контрактом. Скорее всего, они более квалифицированы, имеют хорошее образование, опыт работы, отличаются добросовестностью и не спешат менять одно место работы на другое. Если сравнивать заработки в этих двух группах в период коронавируса, то не удастся установить эффект удаленной работы в период COVID-19. Две группы отличаются как наблюдаемыми характеристиками, которые можно учесть в регрессионной модели для зарплаты, так и ненаблюдаемыми, в частности, добросовестностью, повлиявшей на вероятность попадания в группу воздействия.

Можно попытаться сравнить результирующий показатель в некоторой группе *до и после воздействия*. В данном примере — заработки неформально занятых работников в 2019 и в 2020 гг. Но и это сравнение также будет страдать от искажения искомого эффекта другими шоками, которые произошли в период действия коронавирусных ограничений. Например, в тот же период на мировом рынке сильно упали цены на нефть, которые влияют на экономическую активность, потребительские цены, потребительский спрос, а следовательно, на предложение товаров и услуг и, в конечном счете, на заработки в стране в целом.

Таким образом, ни сравнение двух групп (with–without treatment), ни сравнение двух периодов (before–after treatment) по отдельности не ведет к достоверной оценке влияния регулирующего воздействия — коронавирусных ограничений. Метод DD комбинирует два периода и две группы, тем самым устраняя ненаблюдаемую неоднородность в предположении, что эта неоднородность постоянна во времени (time-invariant heterogeneity). Аналитически модель выглядит следующим образом:

$$Y_{it} = \alpha_i + \gamma_t + \beta D_{it} + \nu_{it}. \quad (1)$$

Здесь Y_{it} — это исследуемый показатель, например, логарифм зарплаты. Сконструируем переменную $D_{it} = D_i T_t$ следующим образом. $D_i = 1$ для индивида i $(i = 1,...,N)$ из группы воздействия, и $D_i = 0$ для контрольной группы. Если это период до воздействия, то $T_t = 0$, если после, то $T_t = 1$. Предположим, что имеются сбалансированные панельные данные, т.е. каждый индивид наблюдался как при $t = 0$, так и при $t = 1$. Сбалансированность панельных данных не критична. Можно анализировать несбалансированные панельные данные или повторяющиеся перекрестные, кросс-секционные данные, например, по двум переписям населения. Ошибка регрессии предполагается состоящей из постоянной во времени для

[3] Учителя могли вести занятия онлайн, но многие были не удовлетворены качеством процесса. Родители, чьи дети не могли посещать детский сад или школу, испытывали двойную нагрузку, работая из дома онлайн и занимаясь одновременно с этим воспитанием и уходом за детьми.

[4] Например, ИТ-специалисты выиграли от экономии на транспортных издержках, работая удаленно.

индивида i ошибки α_i и переменной части ошибки ν_{it}, не коррелирующей с T_s и D_{is} $(s = 0,1)$. Коэффициент γ_t отражает вклад периода воздействия, общий для групп, в результирующий показатель. Постоянная часть ошибки α_i предполагается коррелирующей с D_i, поскольку отбор в группы не случаен. Коэффициенты β, γ состоятельно оцениваются методом фиксированных эффектов (two-way fixed effects, TWFE), поскольку ненаблюдаемая, но постоянная часть ошибки α_i устраняется. Оценки TWFE и по модели в первых разностях (FD, first differences) идентичны. Оценка коэффициента β, которую обозначим β_{FE}[5], в данном случае показывает влияние регулирующего воздействия на группу воздействия. Таким образом, даже учитывая неслучайный отбор в группы *Treated* и *Control*, коэффициент β состоятельно оценивает усредненный эффект воздействия на группу воздействия (the average treatment effect for the treated, *ATT*) в сделанных предположениях.

Измерение эффекта регулирующего воздействия. Подведем к изложенному выше теоретическую базу. Эффект воздействия понимается как различие в результирующей переменной Y_i благодаря воздействию (каким-то политическим мерам, реформам, программам и т.д.) $D_i = 1$ в сравнении с Y_i без воздействия $D_i = 0$:

$$ATE = \mathrm{E}\big(Y_i(1)\big) - \mathrm{E}\big(Y_i(0)\big). \tag{2}$$

Эту оценку называют средним эффектом воздействия (the average treatment effect, *ATE*). В скобках указано состояние — с воздействием и без воздействия. Зная оба исхода (без воздействия и с воздействием) для каждого индивида, можно считать изменение *ATE* следствием воздействия. Проблема измерения ATE состоит в том, что индивиды наблюдаются только в одном состоянии, и неизвестно, каким был бы Y_i без воздействия. Это — фундаментальная проблема оценки причинно-следственной связи между результатом и воздействием, на которую первым обратил внимание Rubin (1974). Выход из положения состоит в дополнении выборки контрольной группой, результирующая переменная в которой, предположительно, служит хорошим прокси для неизвестной части или *контрафактуальным* значением результирующего показателя. Если бы индивиды в контрольной группе и группе воздействия были абсолютно идентичными, то достаточно было бы сравнить средние значения Y по группам. Чаще встречается такая ситуация, что индивиды различны не только по наблюдаемым, но и по ненаблюдаемым характеристикам. Последние могут объяснять более вероятное (не)попадание в группу воздействия. Если ненаблюдаемые характеристики, которые в регрессионной модели попадают в ошибку регрессии, коррелируют с D_i, то возникает проблема эндогенности. Как показано выше, в модели DD проблема эндогенности устраняется применением FE или FD. Сосредоточимся на том, какой эффект измеряет коэффициент β.

В модели DD есть два периода. Группа воздействия получает это воздействие во втором периоде, тогда как в первом периоде обе группы не получают воздействия. Итоговая оценка представляет собой эффект воздействия на индивидов из группы воздействия во второй период. Оценка модели позволяет найти средний эффект воздействия на группу воздействия *ATT*. Здесь для расчета ненаблюдаемого $\mathrm{E}\big(Y_{i1}(0) \mid D_i = 1\big)$ используется условие параллельного тренда (parallel trend assumption, PTA) — предположение о том, что динамика результирующего показателя в группе воздействия без воздействия была бы такой же, как и в контрольной группе (Roth et al., 2023, p. 5):

[5] В данном случае неважно, что оценка была получена методом TWFE, а не FD.

$$ATT = \mathrm{E}\left(Y_{i1}(1) - Y_{i1}(0) \mid D_i = 1\right), \text{ где}$$
$$\mathrm{E}\left(Y_{i1}(0) \mid D_i = 1\right) = \mathrm{E}\left(Y_{i0}(1) \mid D_i = 1\right) + \mathrm{E}\left(Y_{i1}(0) - Y_{i0}(0) \mid D_i = 0\right) = \quad (3)$$
$$= \mathrm{E}\left(Y_{i0} \mid D_i = 1\right) + \mathrm{E}\left(Y_{i1} - Y_{i0} \mid D_i = 0\right).$$

Таким образом,

$$\beta_{FE} = ATT = \mathrm{E}\left(Y_{i1} - Y_{i0} \mid D_i = 1\right) - \mathrm{E}\left(Y_{i1} - Y_{i0} \mid D_i = 0\right). \quad (4)$$

В классической DD модели делается три существенных предположения: 1) взаимная независимость сравниваемых групп (independent sampling или the stable unit treatment value assumption, SUTVA), 2) наличие параллельного тренда (parallel trend assumption, PTA) и 3) неожиданность воздействия (no anticipation period assumption). Обратим внимание, что состояний только два — оказаться под воздействием или нет. Следовательно, переменной воздействия достаточно сопоставить бинарную переменную D_i.

Взаимная независимость сравниваемых групп понимается как то, что результирующий показатель для индивида *i* не связан с тем, оказывалось ли воздействие на индивида *j* (Roth et al., 2023, p. 4). Это довольно сильное предположение, и в данном примере оно не выполняется в долгосрочном периоде, поскольку снижение заработков ведет к снижению спроса на продукцию. Снижение спроса косвенно влияет на спрос на рынке труда как для тех, кто отнесен к группе *Treated*, так и для тех, кто находится в группе сравнения. Следовательно, увольнение одних работников опосредованно ведет к снижению заработков оставшихся. Но в краткосрочном периоде можно не думать о возникновении так называемого spillover эффекта (косвенного воздействия на индивидов в контрольной группе).

Как видно из примера, для проверки выполнения условия SUTVA необходимо понимать те механизмы и каналы, через которые воздействие может повлиять не только на группу *Treated*, но и на контрольную группу. Иногда это допущение можно протестировать. Например, если имеется возможность собрать данные по сравниваемым группам в местности с воздействием (выборка 1) и в географически удаленных местностях, где нет воздействия (выборка 2), то можно сравнить *ATT*, полученный только по выборке 1, с *ATT*, где индивиды группы *Treated* из выборки 1 сравниваются с контрольной группой из выборки 2. В предположении, что выборки 1 и 2 однородны и нет причин ожидать различий в результирующем показателе кроме как из-за воздействия, схожесть или различие оценок двух *ATT* покажет выполнение условия SUTVA в выборке 1.

Следующим условием для состоятельности оценки *ATT* через коэффициент β_{FE} является предположение о том, что ненаблюдаемая неоднородность между индивидами контрольной группы и группы воздействия не меняется во времени. Это условие называется допущением о параллельном тренде. Согласно PTA, при отсутствии воздействия различие между группами по величине *Y* оставалось бы неизменным, следовательно, эффект изменения результирующего показателя происходит исключительно за счет регулирующего воздействия. Недостаток модели 2×2 — это невозможность протестировать PTA. Его можно только обосновать, основываясь на понимании механизмов изменения *Y* в двух группах.

Еще одно важное условие — это неожиданность воздействия (no anticipation effect). Предполагается, что для всех индивидов воздействие оказывается в равной степени неожиданным, что исключает подстройку поведения под ожидаемое воздействие.

3. Несколько групп и периодов воздействия

Конструкция 2×2 интуитивно понятна, чем и объясняется ее популярность для измерения эффекта от регулирующего воздействия. Возникает вопрос, можно ли использовать TWFE для измерения эффекта, когда периодов или сравниваемых групп больше, чем два. Модель (1) легко обобщается на случай нескольких периодов $t = 1,...,T$ и групп $g = 0,...,G$, объединяющий индивидов, которые оказались под воздействием в определенный момент времени.

Рассмотрим примеры. Если анализировать изменение зарплаты в период коронавирусных ограничений не по годам, а по месяцам, то периодов воздействия в 2020 г. становится 9 (апрель—декабрь 2020). Логично уравновесить протяженный период воздействия не менее протяженным периодом до введения ограничений, но правильней будет настроить данные на более длинный период до воздействия, чем после него. Если же учесть региональные ограничения, то периоды воздействия оказываются разными по регионам с учетом разных времен выхода из ограничений. Оценка β_{FE} теперь фактически складывается как некоторым образом средневзвешенное из частных эффектов для отдельных периодов воздействия TE_t, а с учетом различия ограничений по регионам — из TE_{gt}. Каждое g соответствует группе регионов, где ограничения вводились и отменялись в одни и те же моменты времени. Возникает вопрос, соответствует ли полученная оценка β_{FE} эффекту воздействия *ATT*. В общем случае ответ — нет.

Следует отметить, что индекс g во многих статьях по этой теме присваивается не случайным перечислением групп, а связывается с началом воздействия (Callaway, Sant'Anna, 2021). Для примера предположим, что ограничения в связи с коронавирусом вводились постепенно в разных регионах. Пусть в Москве они были введены в апреле 2020 г., а в Удмуртии и Татарстане — в мае 2020 г. Если взять в качестве периода месяц и начать наблюдения, скажем, в октябре 2019 г., то $g = 7$ для Москвы и $g = 8$ для двух других регионов.

Логично ожидать, что даже при сохранении неизменного воздействия может происходить либо накопление эффекта, либо его угасание, либо какие-то колебания. Иными словами, сила воздействия может зависеть от того, сколько периодов p прошло с момента воздействия, поэтому появляется необходимость оценить TE_{gt}. Соответственно, по каждому из индексов может быть произведено усреднение. Объединение индивидов в группы по начальному моменту воздействия позволяет определить средние групповые эффекты TE_{gt} и использовать их для расчета среднего эффекта на определенный момент воздействия *ATT*. Этот эффект полезен для оценки результатов воздействия на экономику в какой-то момент времени. Средний эффект по группе ATT_g, получившей воздействие в определенный момент, полезен для оценки воздействия по группам регионов. ATT_p, усредненный по некоторому периоду воздействия p, будет показывать средний эффект спустя p периодов после начала воздействия. Такой подход носит название event-study. Средневзвешенное групповых эффектов дает *ATT*. Расчет весов зависит от сделанных допущений и выбора метода, которые обсуждаются ниже.

Когда есть несколько групп с варьирующимся началом воздействия, проблема оценки *ATT* проявляется в том, что какие-то группы могут только ожидать воздействия (*not-yet-treated*, *NYT*), а другие — уже находиться под воздействием длительное время. По сравнению с группой, первой испытавшей воздействие, последующие оказываются более готовыми, и могут заранее что-то изменить, чтобы не испытать негативного влияния воздействия

во всей силе. Это, очевидно, усложняет подход к оценке. Возникает вопрос, как сформировать контрольную группу, следует ли в нее включать только те объекты, которые никогда не окажутся под воздействием (*never treated*, *NT*) или же в контрольную группу следует добавить и объекты из *NYT*. И если включать группы, испытывающие воздействие с запозданием, то с каким весом следует учитывать их в расчете контрафактуального $Y(0)$? В общем случае модель (1) трансформируется в модель (5), предполагающую, что воздействие β_{gp} специфично как для группы g, так и для периода p относительно начала воздействия для этой группы. Причем сам показатель и случайная составляющая специфичны для индивида i из группы g и момента t:

$$Y_{gpit} = \alpha_g + \gamma_p + \beta_{gp} D_{gp} + \nu_{gpit}. \quad (5)$$

Следовательно, эффект ATT представляет собой взвешенную сумму отдельных эффектов β_{gp}. В общем случае само воздействие может различаться по силе, т.е. измеряться не бинарно, а некоторым набором значений (dose-response treatment) или непрерывной переменной (continuous treatment). Возникает вопрос в состоятельности оценки ATT методом TWFE.

Когда TWFE оценка состоятельна?

Статический TWFE. В случае нескольких периодов воздействия можно ожидать, что эффект не меняется со временем и одинаков для всех групп воздействия, т.е. необходимо оценить лишь один параметр β_{FE}. В этом случае мы имеем дело со статической моделью. Будет ли β_{FE} состоятельной оценкой ATT для произвольного числа периодов и групп? Roth, Sant'Anna (2021, pp. 5–6) и De Chaisemartin, D'Haultfoeuille (2022) показали, что β_{FE} остается состоятельной для ATT, если (дополнительно к условиям для случая 2×2) выполняются условия: 1) воздействие начинается для всех групп одновременно; и 2) воздействие не прекращается, т.е. однажды попав в группу воздействия, индивид остается в ней до последнего периода T (staggered treatment).

Несмотря на состоятельность самих оценок, их стандартные ошибки будут смещенными из-за потенциальной автокорреляции. Bertrand et al. (2004) обращают внимание на то, что из-за автокорреляции ошибок нулевая гипотеза об отсутствии эффекта воздействия часто будет отвергаться даже тогда, когда эффекта нет. Стандартные ошибки коэффициентов регрессии, кластеризованные по группам — это одно из предложенных ими решений проблемы. San и Abraham (2021) предлагают использовать команду *fixest* для R (*eventstudyinteract* для Stata), где ошибки могут быть скорректированы несколькими способами.

В общем случае, когда воздействие на разные группы начинается в разные моменты времени и оно неодинаково по группам, оценка β_{FE} в модели (1) будет смещенной и несостоятельной. Доказательства этого факта в несколько разном изложении приводятся в работах (Borusyak, Jaravel, 2018; De Chaisemartin, D'Haultfoeuille, 2020; Gardner, 2021; Goodman-Bacon, 2021). Авторы показывают, что оценка β_{FE} может быть представлена как средневзвешенное частных эффектов TE_{gt}, а веса могут быть как положительными, так и отрицательными. Следовательно, даже при всех положительных частных TE_{gt} знак β_{FE} может оказаться отрицательным, и наоборот. Проблема возникает потому, что индивиды из групп, получающих воздействие позже, оказываются в контрольной группе для тех, что получили воздействие раньше. Потом эти же индивиды переходят в группу воздействия. Чем дольше индивиды находятся в той или иной группе, тем больше оказывается вес, с которым их значения Y_{it} учитываются в расчете эффекта. Goodman-Bacon (2021) демонстрирует, как меняются

веса в зависимости от времени начала воздействия. Увидеть значения весов, с которыми TWFE метод учитывает TE_{gp} при расчете β_{FE}, можно командой *eventstudyweights* для Stata[6].

Динамический TWFE. Как сказано выше, при рассмотрении нескольких периодов и групп воздействия можно ожидать, что эффект воздействия будет различаться по периодам от начала воздействия (ATT_p), календарному времени (ATT_t) и группам (ATT_g). И эти оценки представляют отдельный интерес. Очевидно, что усреднение по каждому из этих параметров необязательно ведет к одному и тому же значению *ATT*.

Goodman-Bacon (2021) показывает, что TWFE по-прежнему дает состоятельные оценки и проблема отрицательных весов исчезает, если выполняются следующие условия: 1) усредненное по отдельному периоду воздействие ATT_t не меняется во времени, хотя и может быть различным для групп; 2) воздействие бинарное; 3) нет таких групп, на которые воздействие оказывается большую часть времени; 4) нет таких периодов, когда большая часть групп оказывается под воздействием. Поэтому, чтобы избежать проблемы отрицательных весов, Jakiela (2021) предлагает опустить некоторое число последних периодов и исключить группы, которые почти все время оказываются под воздействием.

Тестирование PTA. Следует отметить, что наличие большого числа периодов до воздействия служит естественной базой для тестирования PTA. Если это условие выполняется, ATT_t до начала воздействия должны быть совместно незначимы. Но даже соблюдение PTA не гарантирует, что *ATT* измеряет причинно-следственную связь между воздействием и результирующим показателем. Kahn-Lang, Lang (2020) приводят следующий пример. Известно, что средний рост девочек и мальчиков до 13 лет примерно одинаков. Но то, что в последующие годы девочки становятся гораздо выше мальчиков, не следует связывать с обрядом бар мицвы. Пример показывает, что понимание природы данных и формирования выборки играет немалую роль в интерпретации результатов.

В принятии и отрицании наличия PTA возникают статистические ошибки I и II рода. Bilinski, Hatfield (2018) указывают на то, что если гипотеза о соблюдении PTA тестируется со значимостью 5%, то вероятность не идентифицировать нарушение PTA (ошибка II рода) может быть значительно больше. Roth et al. (2023) отмечают, что в общем случае любое монотонное функциональное преобразование Y может нарушить PTA, и наоборот, если PTA не соблюдается для Y, оно может соблюдаться для $\ln(Y)$ или другого монотонного преобразования. Методы, предложенные Callaway, Sant'Anna (2021) и Sun, Abraham (2021), позволяют существенно ослабить PTA до PTA только для групп воздействия и только после воздействия. Анализу чувствительности данных к нарушению PTA и природе этого нарушения посвящены работы (Bilinski, Hatfield, 2018; Dette, Schumann, 2020; Freyaldenhoven et al., 2021; Keele et al., 2019; Manski, Pepper, 2018; Rambachan, Roth, 2023; Roth, 2022; Ye et al., 2021), в которых можно найти подробное описание проблемы и подходов к ее решению.

4. Развитие модели

Roth et al. (2023) и De Chaisemartin, D'Haultfoeuille (2022) дают подробный обзор, при каких условиях какая модификация метода разности разностей может обеспечить состоятельные оценки. В частности, Roth et al. (2023) начинают статью своего рода алгоритмом

[6] См. https://github.com/lsun20/EventStudyWeights.

проверки сделанных допущений. Это говорит о том, насколько важны принимаемые исследователем предположения и их соответствие процессу, генерирующему данные, для получения состоятельной оценки *ATT*.

Перечисленные в этом разделе методы исходят из предположения (в дополнение к уже сделанным), что различие между группами состоит во времени воздействия, но не в самом эффекте воздействия, который может быть как статическим, так и динамическим. Иными словами, воздействие однородно по группам, но может быть неоднородно по времени воздействия. Borusyak et al. (2021), Freyaldenhoven et al. (2021), Gardner (2021) и Sun, Abraham (2021) развивают TWFE в двухшаговую процедуру.

Freyaldenhoven et al. (2021) акцентируют внимание на том, что тренд или изменение объясняющих переменных могут исказить результаты тестирования РТА. Авторы расширяют возможности применения TWFE для оценки *ATT*, когда изменение результирующего показателя во времени происходит благодаря воздействию и влиянию некоторого временно́го тренда, не обязательно линейного, но общего для групп воздействия и контрольной группы, или благодаря некоторому изменению объясняющих переменных (confounding factors), влияющих на результирующий показатель. Авторы рекомендуют включить в регрессионную модель тренд и его произведение на D_i. Используя имитационное моделирование, они демонстрируют важность анализа процесса, генерирующего данные. Показывается, что динамика изменения некоторой объясняющей переменной до и после воздействия может существенно исказить оценку *ATT*, т.е. привести к так называемому Ashenfelter's dip (Ashenfelter, 1978), когда кажущееся увеличение эффекта связано на самом деле с тем, что какие-то неучтенные процессы совпали по времени с периодом воздействия для одной из групп. Пакет *xtevent* для Stata, предложенный Freyaldenhoven et al. (2021), позволяет учесть динамику в сравниваемых группах, дополнительно используя отражающие ее переменные в качестве инструментов. Команда *xteventplot* графически иллюстрирует результаты оценивания, а *xteventtest* позволяет протестировать РТА и совместное равенство ATT_t нулю.

Gardner (2021) предлагает использовать двухшаговый метод оценки с использованием TWFE, когда на первом шаге оценивается модель только на фиксированные эффекты в подвыборке, где воздействия не было, т.е. в контрольной группе (*NT*) и группах воздействия до воздействия (*NYT*):

$$Y_{gpit} = \alpha_g + \gamma_p + v_{it}. \tag{6}$$

На втором шаге *ATT* оценивается в сравнении значений *Y* после воздействия со значениями в группах *NT* и *NYT*, элиминируя рассчитанные на первом шаге фиксированные эффекты:

$$Y_{gpit} - \hat{\alpha}_g - \hat{\gamma}_p = \beta_{gp} D_{gp} + u_{gpit}. \tag{7}$$

Gardner (2021) доказывает состоятельность и несмещенность оценки, даже когда воздействие может быть разным по группам и периодам. Метод реализован в пакете R *did2s* (Butts, 2021).

Borusyak et al. (2021) развивают TWFE следующим образом. На первом шаге фиксированные эффекты оцениваются не для периодов до/после воздействия и групп, как рекомендует Gardner (2021), а для индивидов *i* и календарного времени *t* в контрольной группе (*NT*) и группах воздействия до воздействия (*NYT*). Затем предсказанные по регрессии значения используются как контрафактуальные, и разность между Y_{it} для объектов, получивших

воздействие, и предсказанным значением $Y_{it}(0)$ используется как оценка TE_{it}. В зависимости от задач пользователя из полученных оценок рассчитывается ATT, ATT_t, ATT_g или ATT_p как взвешенная сумма полученных оценок TE_{it}. Для расчета ATT достаточно поделить сумму всех эффектов на число наблюдений, на которые оказывалось воздействие. Таким образом, результирующие оценки дают возможность учесть неоднородность эффектов воздействия как по группам, так и по периодам. Авторы демонстрируют, что метод обеспечивает состоятельные оценки, если для всех групп и периодов выполняется PTA, воздействие не прекращается, если уже началось (staggered treatment), и нет ожидания воздействия. Метод реализуется разработанными этими авторами пакетами *did_imputation* для Stata и *didimputation* для R. Он позволяет включать объясняющие переменные, оценивать тройную разность разностей, а также оценивать небинарное воздействие.

Sun и Abraham (2021) предлагают на первом шаге использовать TWFE для оценки TE_{gp}. В выборку могут быть включены все наблюдения, если имеется группа *NT*, или часть наблюдений без соответствующих последним периодам так, чтобы группа, получившая воздействие последней, оставалась в статусе *NYT*. На втором шаге рассчитываются веса, с которыми TE_{gp} входят в финальную оценку ATT, которые равны доле наблюдений группы g в период p. На последнем шаге рассчитывается взвешенное ATT. Авторы демонстрируют, что полученная оценка состоятельна, если соблюдаются условия PTA и нет ожидания воздействия, даже когда воздействие неоднородно по группам. Freyaldenhoven et al. (2021) указывают на то, что в случае эффекта, неоднородного по времени и группам, веса в оценке Sun, Abraham (2021) могут оказаться отрицательными.

Borusyak, Jaravel (2018) и Sun, Abraham (2021) демонстрируют, что динамическая модель (5) позволяет установить причинно-следственную связь между воздействием и изменением результирующего показателя, если неоднородность связана только с разным по времени началом воздействия, но не с характером воздействия на группы. Таким образом, помимо проверки наличия отрицательных весов, в расчете ATT возникает необходимость дополнительно тестировать постоянство воздействия и его однородность по периодам для разных групп, как предлагает Jakiela (2021).

5. Методы оценки альтернативные TWFE

Решением проблемы оценки эффекта, когда воздействие в разных группах неоднородно не только по времени, но также по периодам от начала воздействия и силе воздействия, является использование альтернативных методов. Рассмотрим непараметрические и полупараметрические методы DD оценки, когда число периодов и групп произвольно.

Оценка статического эффекта. De Chaisemartin и D'Haultfoeuille (2020) для статической модели предлагают в качестве оценки ATT использовать усредненные DD в начале и конце воздействия (DD_M). В отличие от методов, предложенных выше, этот метод не требует условия последовательного воздействия, но исходит из предположения, что после начала воздействия эффект остается постоянным до его окончания. В этих предположениях авторы предлагают измерять разность разностей в конструкции 2×2 для групп на пограничных позициях. Таких пограничных позиций две — попадание в воздействие по сравнению с предыдущим периодом (switch-in), когда измеряется $DD_{+,t}$, и выход из воздействия по сравнению с предыдущим периодом воздействия (switch-out), когда измеряется $DD_{-,t}$. Индивиды,

которые не испытывали воздействия ни в один из периодов, служат контрольной группой. Усреднение всех полученных $DD_{+,t}$ и $DD_{-,t}$ дает результирующий DD_M. Авторы демонстрируют состоятельность оценки.

De Chaisemartin, D'Haultfoeuille (2021) показывают, что результат воздействия может быть несмещенно и робастно оценен в этом подходе, даже если эффект неоднороден по группам и различается во времени, может быть небинарным и прекращаться. Подход также позволяет измерить эффект воздействия, если все группы рано или поздно оказываются под воздействием. Процедура реализована как *did_multiplegt* для Stata и R.

Оценка меняющегося, но не прекращающегося во времени эффекта. Если воздействие меняется во времени, например, со временем накапливается негативное влияние эпидемических ограничений на заработки, то значение ATT_t также представляет интерес, как и результирующий *ATT*. Callaway, Sant'Anna (2021) предлагают измерять ATT_t в DD конструкции как средневзвешенное от TE_{gt}, а *ATT* — как средневзвешенное ATT_t тремя методами оценки DD: 1) основанной на регрессионной модели (outcome regression, OR); 2) основанной на методе обратного вероятностного взвешивания (inverse probability weighting, IPW); и 3) комбинацией первого и второго подходов, дающей оценку с двойной устойчивостью (doubly-robust estimand, DR). Авторы демонстрируют, что предложенный ими метод DR позволяет избежать проблемы отрицательных весов. Также метод в явном виде демонстрирует, какие из объектов, не получивших воздействие, были отобраны в контрольную группу.

В предложенных подходах Callaway, Sant'Anna (2021) ослабляют классические предположения, меняя безусловное PTA на условное, которое соблюдается только после учета наблюдаемых характеристик. Авторы рассматривают два условных PTA для сравнения с группами *NT* и *NYT* по отдельности. *ATT* рассчитывается двумя способами в зависимости от сделанных предположений. Авторы также меняют предположение внезапного воздействия на то, что для всех групп воздействие оказывается одинаково неожиданным. Иными словами, предположение состоит в одинаковом по длительности периоде ожидания воздействия (limited treatment anticipation, LTA). Callaway, Sant'Anna (2021) предлагают сдвигать начало воздействия на число периодов, соответствующее «ожиданию». Если, например, начало воздействия оказывается неожиданным для всех групп, то длительность ожидания равна 0. Если за один период до начала воздействия группы в ожидании воздействия меняют свое поведение, и это демонстрирует результирующий показатель *Y*, то длительность ожидания равна 1 и т.д.

Метод IPW (следовательно, и DR) требует дополнительного условия на то, что каждому индивиду в группе воздействия найдутся подходящие индивиды из контрольной группы, имеющие такие же шансы попасть в группу воздействия. Это — так называемое условие перекрывающихся выборок (overlap assumption), которое формально записывается как

$$0 < \Pr\left(D_i = 1 \mid X_i\right) < 1. \tag{8}$$

Смысл этого условия состоит в том, что в выборке нет индивидов, для которых вероятность попасть в группу воздействия равна 0 или 1.

Callaway, Sant'Anna (2021) более подробно рассматривают DR, который по сравнению с OR и IPW обладает двойной устойчивостью. Если для OR важна корректная спецификация модели для *Y* (а следовательно, выполнение условного PTA), то для IPW корректной спецификации модели не требуется, но необходимо корректно специфицировать модель вероятности попадания в группу воздействия. DR допускает, что даже если в спецификации модели

для результирующего показателя или вероятностной модели допущена неточность, метод, тем не менее, обеспечивает состоятельную оценку эффекта воздействия. Оценки ATT_t рассчитываются для одного из вариантов группы сравнения, *NT*, *NYT* или обеих групп, в зависимости от сделанного исследователем предположения, для какой именно группы выполняется условное PTA. Стандартные ошибки рассчитываются дельта-методом или бутстраппированием. Метод реализован авторами как пакеты *csdid* для Stata и *did* для R.

Roth, Sant'Anna (2021) предлагают plug-in оценку, объединяющую подходы (Callaway, Sant'Anna, 2021; De Chaisemartin, D'Haultfoeuille, 2020; Sun, Abraham, 2021) и TWFE. Основное предположение, которое делают Roth и Sant'Anna, состоит в том, что время воздействия случайно или псевдослучайно (что сильнее, чем PTA), а воздействие происходит последовательно. Авторы определяют эффект воздействия для индивида *i* в момент времени *t* как $TE_{it,gg'} = Y_{it}(g) - Y_{it}(g')$ — различие в результирующей переменной, если бы воздействие начиналось в момент *g*, а не в момент g'. Средневзвешенное этих эффектов по выборке, из которого вычитается различие результирующей переменной (реальной или предсказанной) в сравниваемых группах до момента воздействия, дает *ATE*. Расчет весов здесь опускается. Авторы, используя метод Монте-Карло, демонстрируют хорошие свойства предложенного подхода, в частности, наименьшую стандартную ошибку оценки.

Сравнение методов. Можно провести параллель между методами Callaway, Sant'Anna (2021) и De Chaisemartin, D'Haultfoeuille (2020). Последний не требует последовательного воздействия. Roth et al. (2023) указывают, что метод De Chaisemartin, D'Haultfoeuille (2020) соответствует оценке $ATT_{t=1}$, предложенной Callaway, Sant'Anna (2021) для одного периода вперед.

Borusyak et al. (2021), Freyaldenhoven et al. (2021), Gardner (2021) и Sun, Abraham (2021) основывают сравнение на средних результирующего показателя до воздействия. DR метод, предложенный Callaway, Sant'Anna (2021), и метод De Chaisemartin, D'Haultfoeuille (2020) строят сравнение групп относительно последнего периода перед воздействием. Выполнение PTA в предложенных Borusyak et al. (2021), Freyaldenhoven et al. (2021), Gardner (2021) и Sun, Abraham (2021) методах требуется для всех периодов до воздействия и для всех групп, а в методе Callaway, Sant'Anna (2021) и De Chaisemartin, D'Haultfoeuille (2020) достаточно двух периодов. Roth et al. (2023) отмечают, что более полный учет неоднородности разной природы, как правило, ведет к потере эффективности оценок.

Borusyak et al. (2021) показывают, что полученные ими оценки будут более аккуратными (т.е. имеющими более узкий доверительный интервал) в сравнении с предложенными Callaway, Sant'Anna (2021) и Sun, Abraham (2021), если ошибки регрессии гомоскедастичны и не автокоррелированы. Авторы также демонстрируют устойчивость результатов даже при наличии автокорреляции.

Квантильный эффект воздействия. Можно предположить, что негативный эффект воздействия COVID-19 на заработки наблюдается в наиболее бедных 25- или 10%-х доходных группах и не наблюдается для работника со средним заработком. Athey, Imbens (2006) предлагают измерять изменения в изменениях (Changes-in-Changes, DD-CIC), основанные на оценке распределения $Y(0)$ для группы воздействия. Оценка строится на предположении, что соотношение распределений сравниваемых групп неизменно во времени, т.е. если 20%-ный квантиль распределения одной группы соответствовал 45%-ному квантилю другой в начальный момент времени, то это соотношение сохранится и в следующий период, если не будет воздействия. Развитие этой темы и несколько иные предположения для квантильной DD оценки можно найти в (Bonhomme, Sauder, 2011; Callaway, Li, 2019; Roth, Sant'Anna, 2023).

Реализацию квантильной оценки для двух периодов в Stata предлагают пакеты *fuzzydid* (De Chaisemartin, D'Haultfoeuille, 2018, 2020), *rifhdreg* (Rios-Avila, 2020), а также *diff* (Villa, 2016). Последний реализует оценку квантильного эффекта в модели разности разностей с мэтчингом на первом шаге, как это описано в (Meyer, Viscusi, 1995; Heckman et al., 1998).

Вероятное воздействие. В силу различных обстоятельств индивиды из контрольной группы могут оказаться в числе попавших под воздействие, а индивиды из группы воздействия — его избежать. Например, при коронавирусных ограничениях на заработки региональные меры не действовали на тех, кто работал вахтовым методом в других регионах. Таким образом, вероятность попасть в группу воздействия для вахтовиков зависела от мер, принимаемых в регионе занятости, о которых ничего не известно. Иными словами, вероятность воздействия не обязательно равна 0 или 1 (fuzzy DD). Таким образом, $D_{gt} = d$ — некоторой упорядоченной или непрерывной переменной.

Для ситуации с вероятным воздействием, когда воздействие различно по группам или по времени (continuous treatment, dose-response treatment), а также когда индивиды могут попадать и покидать воздействие (non-staggered treatment), De Chaisemartin, D'Haultfoeuille (2018, 2020) предлагают четыре вида оценки:

1) W_{DD} — оценка типа Вальда DD (Wald, 1940);
2) W_{TD} — оценка типа Вальда DD, скорректированная с учетом изменений во времени (the time-corrected Wald ratio);
3) W_{CC} — оценка изменения изменений (the changes-in-changes Wald ratio), предложенная Athey, Imbens (2006);
4) локальный квантильный эффект воздействия (local quantile treatment effect, LQTE).

Последние две связаны друг с другом. В зависимости от того, какие из предположений выполняются, одна из оценок дает *ATT*. Так же как и DD_M, предложенные оценки основаны на изменении Y в момент перехода группы воздействия из одного состояния в другое (начало воздействия или конец воздействия).

Для оценки вероятного воздействия предполагается, что: 1) группа воздействия испытывает бóльший эффект, чем контрольная; 2) доля индивидов, попавших под воздействие в контрольной группе, относительно постоянна во времени; 3) переход от «не под воздействием» в «под воздействием» (или обратно) справедлив для нескольких последующих периодов, но значение, которое принимает (не бинарная) переменная D, наблюдается только в конечный период воздействия (или выхода из него).

W_{DD} представляет собой коэффициент при D_{gt} в двухшаговой оценке регрессии Y на D_{gt} с бинарными переменными T_t (равными 1 в период t и 0 иначе), и переменной G_g для группы, где $G_g T_t$ служит инструментом в регрессии для воздействия D_{gt} (здесь G_g принимает значение 1 в момент начала воздействия и –1 в момент окончания воздействия, в остальные периоды $G_g = 0$). De Chaisemartin, D'Haultfoeuille (2018, 2020) показывают, что W_{DD} состоятельно оценивает *ATT* при условии статического воздействия.

Оценка W_{TC} состоятельно оценивает *ATT* при выполнении условного PTA для всех групп. А оценка W_{CC} состоятельно оценивает *ATT*, если: 1) результирующая переменная Y может быть представлена как монотонно возрастающая функция от постоянных во времени ненаблюдаемых переменных; 2) функция распределения непрерывна и монотонно возрастает для любого Y_{gt} ($D_{gt} = d$).

De Chaisemartin, D'Haultfoeuille (2018) обосновывают применение методов в 2×2 конструкции для fuzzy DD с бинарным воздействием и без объясняющих переменных. Затем

они расширяют предложенную методологию на несколько периодов, небинарное воздействие, включение объясняющих переменных и непоследовательное воздействие. Процедура реализована авторами как *fuzzydid* для Stata и R.

Помимо расчета самих эффектов, авторы пакета дают возможность протестировать равенство оценок по каждой паре. Кроме стандартных тестов на условное и безусловное PTA, они предлагают реализовать тест placebo, искусственно разместив периоды воздействия там, где их на самом деле нет. Сделанные предположения в совокупности со статистическими тестами позволяют выбрать какую-то одну из полученных W_{DD}, W_{TC} и W_{CC} оценок.

Следует отметить, что подходы для оценки W_{DD}, W_{TC} и W_{CC}, предложенные De Chaisemartin, D'Haultfoeuille (2018, 2020), в общем случае используют объясняющие переменные для всех периодов, включенных в оценку. Это отличает их от метода, предложенного Callaway, Sant'Anna (2021), использующего только наблюдения для одного периода перед воздействием и для периодов под воздействием.

6. Имитационное моделирование

Существует немало статей, сравнивающих эффективность и точность предложенных методов. Например, Freyaldenhoven et al. (2021) тестируют четыре процесса, генерирующих данные: статический эффект, эффект с монотонным трендом, эффект при отсутствии пре-тренда и наличии нескольких переменных, влияющих на разную динамику Y в контрольной группе и в группе воздействия. Они тестируют TWFE и методы De Chaisemartin, D'Haultfoeuille (2021), Sun, Abraham (2021) и Borusyak et al. (2021) и делают вывод о том, что новые методы не выигрывают по сравнению с TWFE в величине смещения оценки от медианной.

Ниже приводятся результаты имитационного моделирования, сравнивающего TWFE с его развитием в процедуре *xtevent*, предложенной Freyaldenhoven et al. (2021), и методами Borusyak et al. (2021), Callaway, Sant'Anna (2021), Gardner (2021) и Sun, Abraham (2021), для модели со статическим эффектом, нескольких групп и периодов[7]. Модель, генерирующая данные, представляет собой зависимость Y от X и W, где X также зависит от W, как это нередко случается в экономических процессах. Ошибка кластеризована по индивидам. Процесс, описывающий данные, задан следующим образом:

$$
\begin{aligned}
&Y_{itg} = 3X_{it} + W_{it} + 0.3D_{itg} + \alpha_i + \nu_{it}, \quad X_{it} = 2W_{it} + u_{it},\\
&W_{it}, u_{it}, \nu_{it}, \alpha_i \sim N(0,1), \quad D_{ig} \sim [uniform(0,1)]\ ^{8}, \qquad (9)\\
&\operatorname{cov}(u_{it}, \nu_{js}) = \operatorname{cov}(\nu_{it}, \alpha_j) = \operatorname{cov}(u_{it}, \alpha_j) =\\
&\quad = \operatorname{cov}(D_{itg}, \nu_{js}) = 0, \quad i, j = 1,...,1000, \ t, s = 1,...,10,\\
&1)\ \ G = 6 \ \text{ или } \ 2)\ \ G = \{4,6,9\}.
\end{aligned}
$$

[7] Вычисления можно повторить, используя коды для Stata и R, размещенные в Kotyrlo (2023), "Difference-in-difference estimator simulations: 7 approaches", Mendeley Data, V1, DOI: 10.17632/fd6x8p4xh7.1.

[8] Квадратные скобки в формуле означают округление до ближайшего целого. Таким образом, воздействие задается бинарной переменной.

Оценивается модель для Y_{itg} с целью установить величину эффекта воздействия *ATT*, который в истинной модели равен 0.3. Индекс *itg* в D_{itg} означает, что для индивида *i* в момент времени $t = g$ начинается воздействие. В первой модели — одна группа, испытывающая воздействие, с началом воздействия в $t = 6$, $G = 6$. Во второй модели — три группы, испытывающих воздействие в разные периоды $G_1 = 4$, $G_2 = 6$, $G_3 = 9$, и контрольная группа. Распределение по группам равномерное. Оценка модели осуществляется так, как если бы исследователь предполагал возможную динамику в эффектах, поэтому оцениваемая модель включает временные эффекты воздействия ATT_t или ATT_p в зависимости от применяемого метода. Таким образом, исследователь наблюдает *Y*, *X*, *W*, и *D*, но каков отклик на воздействие — постоянный или меняющийся по величине во времени — ему неизвестно. Для двух сравниваемых групп оцениваются $ATT_{t=6},...,ATT_{t=10}$. Для трех групп воздействия и одной контрольной группы оцениваются ATT_{gp} от начала воздействия на определенную группу. Число симуляций для каждого метода равно 1000.

Таблица 1 представляет усредненный по симуляциям эффект *ATT*, усредненную стандартную ошибку, усредненные *F*- и *p*-value для результатов теста Вальда на равенство временных эффектов воздействия (статический эффект) и равенство временных эффектов до начала воздействия 0 (PTA). В методе Borusyak et al. (2021) (R *didimputation*) и Gardner (2021) (R *did2s*) рассчитывается только *ATT* и стандартная ошибка. Реализация тестов недоступна в (Borusyak et al., 2021). Метод Callaway, Sant'Anna (2021) выдает тест PTA, но не позволяет протестировать равенство ATT_t. Кластеризация ошибок по индивидам в явном виде указывается в TWFE и процедурах (Freyaldenhoven et al., 2021; Gardner, 2021; Sun, Abraham, 2021). В (Borusyak et al., 2021) кластеризация по индивидам предполагается по умолчанию.

По результатам табл. 1 все методы свидетельствуют о том, что PTA не нарушается. Отклонение оценки *ATT* от истинного значения во всех методах не превосходит 5%. Метод, предложенный Borusyak et al. (2021), для одной группы воздействия демонстрирует отклонение на 6%, которое можно считать относительно близким к 5%. Для модели 1 наименьшая стандартная ошибка получена методом TWFE. Freyaldenhoven et al. (2021) и Sun, Abraham (2021) также дают близкие значения стандартной ошибки и значимые оценки *ATT*. Для модели 2 стандартная ошибка минимальна в методе, предложенном Freyaldenhoven et al. (2021). TWFE дает бóльшую стандартную ошибку в сравнении с (Freyaldenhoven et al., 2021) и (Sun, Abraham, 2021) в случае трех групп воздействия. Borusyak et al. (2021) и Callaway, Sant'Anna (2021) дают самые большие стандартные ошибки, и эффект воздействия оказывается незначимым.

Также все методы подтверждают равенство временных эффектов воздействия, несмотря на то что в случае трех групп TWFE рассчитывает оценку по календарным временным эффектам, как в (1), а другие методы — оценку по периодам от начала воздействия, т.е. на основе (5). Таким образом, TWFE тестирует 7 оценок начиная с 4-го периода. В расчете ATT_t для $t = 1$ всего одна группа находится под воздействием, остальные — либо *NT*, либо *NYT*. Начиная с $t = 9$ эффект рассчитывается по трем группам, попавшим под воздействие, в сравнении с группой *NT*. А другие методы, позволяющие анализировать воздействие от начала события (event-study), рассчитывают первый эффект для всех трех групп в сравнении с группой *NT*. Начиная с периода $p = 3$ в расчет включаются только две группы, которые сравниваются с *NT* ($G = 4, 6$), а с периода $p = 6$ — только одна группа, в которой воздействие началось раньше остальных, сравнивается с группой *NT* ($G = 4$).

Таблица 1. Результаты имитационного моделирования. Оценки *ATT*, гипотез о выполнении PTA и равенстве ATT_p

	TWFE	Freyaldenhoven et al. (2021)	Sun, Abraham (2021)	Callaway, Sant'Anna (2021)	Borusyak et al. (2021)
Модель 1. Одна группа воздействия					
PTA *F*-value	1	0.937	0.913	3.592	
PTA *p*-value	0.499	0.522	0.529	0.545	
ATT	0.297***	0.299***	0.297***	0.302	0.282
ATT SE	0.069	0.073	0.073	0.225	0.307
Равенство ATT_p *F*-value	0.993	0.89	0.894		
Равенство ATT_p *p*-value	0.505	0.545	0.544		
Модель 2. Три группы воздействия и контрольная группа.					
PTA *F*-value		0.97	0.902	13.335	
PTA *p*-value		0.51	0.533	0.482	
ATT	0.306***	0.302***	0.301***	0.302	0.300
ATT SE	0.091	0.053	0.065	0.261	0.306
Равенство ATT_p *F*-value	1.016	0.884	0.933		
Равенство ATT_p *p*-value	0.493	0.55	0.53		

Примечание. В оценке методом (Callaway, Sant'Anna, 2021) критическое значение усреднено по испытаниям[9]. *** — $p < 0.01$.

Процедура, реализованная пакетом *did2s*, позволяет получить ATT_p и их стандартные ошибки для шести методов оценки: TWFE, (Sun, Abraham, 2021), (Callaway, Sant'Anna, 2020), (Roth, Sant'Anna, 2021), (Borusyak et al., 2021) и (Gardner, 2021) — табл. 2. Поскольку ковариационные матрицы пакетом *did2s* не возвращаются, оценить значимость *ATT* и проверить равенство временных эффектов воздействия и PTA не представляется возможным. Для одной группы *Treated* все методы дают оценки с отклонением от истинного значения не более 5%. Для трех групп методы Roth, Sant'Anna (2021) и Callaway, Sant'Anna (2020) дают существенное отклонение в оценках ATT_p. Наименьшие стандартные ошибки дают методы Borusyak et al. (2021) и Gardner (2021). TWFE и Sun, Abraham (2021) обеспечивают значимость эффекта, тогда как Roth, Sant'Anna (2021) и Callaway, Sant'Anna (2020) не дают значимого эффекта воздействия. Таким образом, в случае статического эффекта воздействия TWFE обеспечивает довольно точную оценку *ATT*, даже когда воздействие оказывается в разные периоды. Однако методы Borusyak et al. (2021), Gardner (2021) и Freyaldenhoven et al. (2021) дают более аккуратные оценки. Идея комбинации IPW и DD в (Callaway, Sant'Anna, 2020) себя не оправдывает, приводит к неточному измерению величины эффекта и увеличивает стандартную ошибку.

[9] Callaway и Sant'Anna (2021) не используют стандартные критические значения и рассчитывают их в процессе оценки.

Таблица 2. Оценки эффекта воздействия по периодам от начала воздействия

Период	*ATT*	*ATT* SE	Период	*ATT*	*ATT* SE
(1)	(2)	(3)	(4)	(5)	(6)
		TWFE			
			–8	0.000	0.100
			–7	–0.003	0.099
			–6	–0.005	0.089
–5	0.002	0.089	–5	0.000	0.075
–4	0.000	0.090	–4	–0.001	0.069
–3	–0.003	0.089	–3	–0.001	0.059
–2	0.004	0.089	–2	–0.001	0.063
0	0.301***	0.089	0	0.297***	0.063
1	0.302***	0.090	1	0.299***	0.060
2	0.303***	0.090	2	0.299***	0.068
3	0.301***	0.089	3	0.299***	0.076
4	0.301***	0.089	4	0.303***	0.072
			5	0.302***	0.100
			6	0.302***	0.099
		Gardner (2021)			
			–8	0.001	0.050
			–7	–0.001	0.050
			–6	–0.003	0.050
–5	0.001	0.028	–5	0.001	0.036
–4	0.000	0.028	–4	0.000	0.036
–3	–0.002	0.028	–3	0.000	0.027
–2	0.002	0.028	–2	0.000	0.026
–1	0.000	0.028	–1	0.001	0.026
0	0.301***	0.069	0	0.297***	0.051
1	0.302***	0.069	1	0.300***	0.051
2	0.303***	0.069	2	0.299***	0.063
3	0.300***	0.069	3	0.299***	0.067
4	0.301***	0.069	4	0.303***	0.067
			5	0.301***	0.101
			6	0.303***	0.101
		Callaway, Sant'Anna (2020)			
			–7	–0.046	0.696
			–6	–0.031	0.696
			–5	0.029	0.695
–4	–0.004	0.491	–4	–0.005	0.491
–3	–0.010	0.490	–3	0.000	0.493
–2	–0.005	0.491	–2	0.003	0.403
–1	–0.010	0.491	–1	0.008	0.402
0	0.318	0.492	0	0.305	0.402

Продолжение табл. 2

Период	*ATT*	*ATT* SE	Период	*ATT*	*ATT* SE
(1)	(2)	(3)	(4)	(5)	(6)
1	0.307	0.491	1	0.302	0.400
2	0.289	0.491	2	0.296	0.492
3	0.303	0.490	3	0.307	0.492
4	0.300	0.492	4	0.305	0.492
			5	0.344	0.695
			6	0.306	0.696
		Sun, Abraham (2021)			
			–8	0.001	0.127
			–7	–0.003	0.127
			–6	–0.004	0.127
–5	0.002	0.089	–5	0.002	0.090
–4	0.000	0.090	–4	–0.001	0.078
–3	–0.003	0.089	–3	–0.001	0.067
–2	0.004	0.089	–2	–0.001	0.073
0	0.301***	0.089	0	0.297***	0.073
1	0.302***	0.090	1	0.300***	0.067
2	0.303***	0.090	2	0.298***	0.089
3	0.301***	0.089	3	0.298***	0.090
4	0.301***	0.089	4	0.303***	0.090
			5	0.299**	0.127
			6	0.302**	0.127
		Roth, Sant'Anna (2021)			
			–7	0.159	0.349
			–6	0.154	0.348
			–5	0.162	0.434
–4	–0.003	0.245	–4	0.156	0.434
–3	–0.006	0.245	–3	0.011	0.187
–2	–0.006	0.245	–2	–0.009	0.184
0	0.315	0.488	0	0.101	0.257
1	0.303	0.489	1	0.105	0.255
2	0.285	0.489	2	0.202	0.336
3	0.298	0.489	3	0.210	0.345
4	0.296	0.489	4	0.282	0.414
			5	0.317	0.498
			6	0.293	0.520
		Borusyak et al. (2021)			
			–8	0.001	0.110
			–7	–0.002	0.108
			–6	–0.005	0.097
–5	–0.001	0.040	–5	0.001	0.080
–4	–0.003	0.075	–4	–0.001	0.073

Окончание табл. 2

Период	*ATT*	*ATT* SE	Период	*ATT*	*ATT* SE
(1)	(2)	(3)	(4)	(5)	(6)
–3	–0.006	0.075	–3	–0.001	0.063
–2	0.001	0.075	–2	–0.001	0.067
–1	–0.001	0.075	0	0.297***	0.051
0	0.301***	0.069	1	0.300***	0.051
1	0.302***	0.069	2	0.299***	0.063
2	0.303***	0.069	3	0.299***	0.067
3	0.300***	0.069	4	0.303***	0.067
4	0.298***	0.069	5	0.301***	0.101
			6	0.303***	0.101

Примечание. Столбцы (1)–(3) соответствуют модели с одним периодом воздействия для всех групп. Столбцы (4)–(6) соответствуют модели с тремя периодами воздействия. Таким образом, в первой группе периоды p меняются от –3 до 6, во второй — от –5 до 4, а в третьей от –8 до 1. *** — $p < 0.01$, ** — $p < 0.05$.

7. Заключение

Статья кратко знакомит читателя с примерами, когда использование TWFE для оценки эффекта воздействия методом разности разностей дает состоятельные и несостоятельные оценки. В общем случае оценки TWFE несостоятельны для нескольких периодов, нескольких групп, начала воздействия, варьирующегося по группам, вероятного воздействия, воздействия, которое прекращается и возобновляется, небинарного и квантильного воздействия. Обоснование несостоятельности приводится, например, в (Borusyak, Jaravel, 2018; De Chaisemartin, D'Haultfoeuille, 2020; Gardner, 2021; Goodman-Bacon, 2021). Эта проблема послужила отправной точкой для развития других подходов к измерению причинно-следственной связи между воздействием и изменением результирующего показателя методом разности разностей. Методы исходят из различных предположений о процессе, порождающем данные, необходимые для получения состоятельной оценки.

В статье рассмотрены различные подходы к измерению эффекта воздействия методом разности разностей. На простом примере с помощью имитационного моделирования показано, что для статического эффекта TWFE дает весьма близкую к истинному значению оценку эффекта, даже если групп несколько и начало воздействия меняется по группам. Однако стандартная ошибка в случае нескольких групп, воздействие на которые начинается в разные периоды, меньше в методах, предложенных Borusyak et al. (2021), Gardner (2021) и Freyaldenhoven et al. (2021). Комбинирование IPW и DD в методе Callaway, Sant'Anna (2020), а также оценка, предложенная Roth, Sant'Anna (2021), дают худшие результаты. В последних двух случаях по некоторым периодам после воздействия оценка далека от истинной, а стандартные ошибки велики. Это доказывает необходимость продолжения исследований по тестированию предложенных методов для различных исходных условий.

Статья не преследовала целью подробно изложить каждый из методов, поскольку они подробно представлены в работах самих авторов. Целью настоящей статьи было указать читателю те важные результаты в применении DD, которые появились в последнее десятилетие,

дать общее представление о развитии метода и о препятствиях к получению достоверных оценок. За рамками рассмотрения осталось применение разности разностей к выборке с малым числом сравниваемых объектов. Важным нерассмотренным аспектом применения метода является эффект воздействия на контрольную группу (spillover effect). Ссылки на необходимую литературу можно найти в последних обзорах по данной теме, например, в (Roth et al., 2023) и других источниках.

Благодарности. Публикация подготовлена в результате проведения исследования по проекту № 23-00-033 «Оценка влияния макрошоков на социально-экономические процессы в регионах России (на примере COVID-19)» в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)». Автор благодарит Андрея Валентиновича Аистова за подробные комментарии и дополнительную информацию, которая помогла сделать статью более полной и точной.

Список литературы

Вулдридж В. М. (2009). Оценивание методом «разность разностей». *Квантиль*, 6, 25–47.

Ashenfelter O. (1978). Estimating the effect of training programs on earnings. *Review of Economics and Statistics*, 60 (1), 47–57.

Ashenfelter O., Card D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics*, 67, 648–660.

Athey S., Imbens G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74 (2), 431–497. DOI: 10.1111/j.1468-0262.2006.00668.x.

Bertrand M., Duflo E., Mullainathan S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119 (1), 249–275. DOI: 10.1162/003355304772839588.

Bilinski A., Hatfield L. A. (2018). Seeking evidence of absence: Reconsidering tests of model assumptions. *ArXiv:*1805.03273. DOI: 10.48550/arXiv.1805.03273.

Bonhomme S., Sauder U. (2011). Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *Review of Economics and Statistics*, 93 (2), 479–494. DOI: 10.1162/REST_a_00164.

Borusyak K., Jaravel X. (2018). Revisiting event study designs with an application to the estimation of the marginal propensity to consume. https://scholar.harvard.edu/files/borusyak/files/borusyak_jaravel_event_studies.pdf.

Borusyak K., Jaravel X., Spiess J. (2021). Revisiting event study designs: Robust and efficient estimation. *ArXiv*:2108.12419. DOI: 10.48550/arXiv.2108.12419.

Butts K. (2021). did2s: Two-stage difference-in-differences following Gardner (2021). R package version 1.0.2. https://cran.r-project.org/web/packages/did2s/vignettes/Two-Stage-Difference-in-Differences.html.

Callaway B., Sant'Anna P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225, 200–230. DOI: 10.1016/j.jeconom.2020.12.001.

Callaway B., Li T. (2019). Quantile treatment effects in difference in differences models with panel data. *Quantitative Economics*, 10 (4), 1579–1618. DOI: 10.3982/QE935.

De Chaisemartin C., D'Haultfoeuille X. (2018). Fuzzy differences-in-differences. *The Review of Economic Studies*, 85 (2), 999–1028. DOI: 10.1093/restud/rdx049.

De Chaisemartin C., D'Haultfoeuille X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110 (9), 2964–2996.

De Chaisemartin C., D'Haultfoeuille X. (2021). Difference-in-differences estimators of intertemporal treatment effects. *NBER Working Paper* No. 29873.

De Chaisemartin C., D'Haultfoeuille X. (2022). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *NBER Working Paper* No. 29691.

Dette H., Schumann M. (2020). Difference-in-differences estimation under non-parallel trends. https://www.ruhr-uni-bochum.de/imperia/md/content/mathematik3/publications/dette_schumann2020.pdf.

Freyaldenhoven S., Hansen C., Pérez J. P., Shapiro J. M. (2021). Visualization, identification, and estimation in the linear panel event-study design. *NBER Working Paper* No. 29170.

Gardner J. (2021). Two-stage differences in differences. https://jrgcmu.github.io/2sdd_current.pdf.

Goodman-Bacon A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225, 254–277. DOI: 10.1016/j.jeconom.2021.03.014.

Heckman J., Ichimura H., Todd P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65 (2), 261–294.

Jakiela P. (2021), Simple diagnostics for two-way fixed effects. *ArXiv*:2103.13229. DOI: 10.48550/arXiv.2103.13229.

Kahn-Lang A., Lang K. (2020). The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business & Economic Statistics*, 38 (3), 613–620. DOI: 10.1080/07350015.2018.1546591.

Keele L. J., Small D. S., Hsu J. Y., Fogarty C. B. (2019). Patterns of effects and sensitivity analysis for differences-in-differences. *ArXiv*:1901.01869. DOI: 10.48550/arXiv.1901.01869.

Manski C. F., Pepper J. V. (2018). How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions. *The Review of Economics and Statistics*, 100 (2), 232–244. DOI: 10.1162/REST_a_00689.

Meyer B., Viscusi W. (1995). Workers' compensation and injury duration: Evidence from a natural experiment. *The American Economic Review*, 85 (3), 322–340.

Rambachan A., Roth J. (2023). A more credible approach to parallel trends. *The Review of Economic Studies*, 90, 2555–2591. DOI:10.1093/restud/rdad018.

Rios-Avila F. (2020). Recentered influence functions (RIFs) in Stata: RIF regression and RIF decomposition. *Stata Journal*, 20 (1), 51–94. DOI: 10.1177/1536867X20909690.

Roth J. (2022). Pre-test with caution: Event-study estimates after testing for parallel trends. *The American Economic Review: Insights*, 4 (3), 305–322. DOI: 10.1257/aeri.20210236.

Roth J., Sant'Anna P. H. (2021). Efficient estimation for staggered rollout designs. *ArXiv:2102.01291*. DOI: 10.48550/arXiv.2102.01291.

Roth J., Sant'Anna P. H., Bilinski A., Poe J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235 (2), 2218–2244. DOI: 10.1016/j.jeconom.2023.03.008.

Roth J., Sant'Anna P. H. (2023). When is parallel trends sensitive to functional form? *Econometrica*, 91 (2), 737–747. DOI: 10.3982/ECTA19402.

Rubin D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66 (5), 670–688.

Sun L., Abraham S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225, 175–199. DOI: 10.1016/j.jeconom.2020.09.006.

Villa J. M. (2016). diff: Simplifying the estimation of difference-in-differences treatment effects. *Stata Journal*, 16, 52–71.

Wald A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*, 11, 284–300.

Ye T., Keele L., Hasegawa R., Small D. S. (2021). A negative correlation strategy for bracketing in difference-in-differences. *ArXiv*: 2006.02423. DOI: 10.48550/arXiv.2006.02423.

Поступила в редакцию 14.03.2023;
принята в печать 02.10.2023.

Kotyrlo E. S. Simple and complex difference-in-differences approach. *Applied Econometrics*, 2024, v. 73, pp. 119–142.

DOI: 10.22394/1993-7601-2024-73-119-142

Elena Kotyrlo
HSE University, Moscow, Russian Federation;
ekotyrlo@hse.ru

Simple and complex difference-in-differences approach

The paper presents extensions of the popular difference-in-differences approach (DD) from 2×2 design on multiple time-period, multiple groups, fuzzy DD, non-staggered treatment and approaches to measure distributional treatment effect. The paper describes assumptions for consistent estimation of the treatment effect by two-way fixed effects model (TWFE) and presents the problem leading to inconsistent estimates justifying the application of alternative estimators. The paper briefly introduces methods developing DD for multiple-period multiple-group cases based on TWFE and alternative approaches. The proposed techniques allow treatment evaluation in the frame of DD when canonical TWFE leads to inconsistent estimates. Some approaches allow replacement of the well-known parallel trend assumption (PTA) for a conditional PTA or time randomisation. The paper refers to implementations of these methods in Stata and R. Simulation modelling demonstrates that the stated properties of the alternative estimators are not always reliable.

Keywords: difference-in-differences; parallel trend assumption; non-staggered treatment; fuzzy DD.

JEL classification: C18; C23; C87.

References

Wooldridge J. M. (2009). Difference-in-differences estimation. *Quantile*, 6, 25–47 (in Russian).

Ashenfelter O. (1978). Estimating the effect of training programs on earnings. *Review of Economics and Statistics*, 60 (1), 47–57.

Ashenfelter O., Card D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics*, 67, 648–660.

Athey S., Imbens G.W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74 (2), 431–497. DOI: 10.1111/j.1468-0262.2006.00668.x.

Bertrand M., Duflo E., Mullainathan S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119 (1), 249–275. DOI: 10.1162/003355304772839588.

Bilinski A., Hatfield L.A. (2018). Seeking evidence of absence: Reconsidering tests of model assumptions. *ArXiv:*1805.03273. DOI: 10.48550/arXiv.1805.03273.

Bonhomme S., Sauder U. (2011). Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *Review of Economics and Statistics*, 93 (2), 479–494. DOI: 10.1162/REST_a_00164.

Borusyak K., Jaravel X. (2018). Revisiting event study designs with an application to the estimation of the marginal propensity to consume. https://scholar.harvard.edu/files/borusyak/files/borusyak_jaravel_event_studies.pdf.

Borusyak K., Jaravel X., Spiess J. (2021). Revisiting event study designs: Robust and efficient estimation. *ArXiv*:2108.12419. DOI: 10.48550/arXiv.2108.12419.

Butts K. (2021). did2s: Two-stage difference-in-differences following Gardner (2021). R package version 1.0.2. https://cran.r-project.org/web/packages/did2s/vignettes/Two-Stage-Difference-in-Differences.html.

Callaway B., Sant'Anna P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225, 200–230. DOI: 10.1016/j.jeconom.2020.12.001.

Callaway B., Li T. (2019). Quantile treatment effects in difference in differences models with panel data. *Quantitative Economics*, 10 (4), 1579–1618. DOI: 10.3982/QE935.

De Chaisemartin C., D'Haultfoeuille X. (2018). Fuzzy differences-in-differences. *The Review of Economic Studies*, 85 (2), 999–1028. DOI: 10.1093/restud/rdx049.

De Chaisemartin C., D'Haultfoeuille X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110 (9), 2964–2996.

De Chaisemartin C., D'Haultfoeuille X. (2021). Difference-in-differences estimators of intertemporal treatment effects. *NBER Working Paper* No. 29873.

De Chaisemartin C., D'Haultfoeuille X. (2022). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *NBER Working Paper* No. 29691.

Dette H., Schumann M. (2020). Difference-in-differences estimation under non-parallel trends. https://www.ruhr-uni-bochum.de/imperia/md/content/mathematik3/publications/dette_schumann2020.pdf.

Freyaldenhoven S., Hansen C., Pérez J.P., Shapiro J.M. (2021). Visualization, identification, and estimation in the linear panel event-study design. *NBER Working Paper* No. 29170.

Gardner J. (2021). Two-stage differences in differences. https://jrgcmu.github.io/2sdd_current.pdf.

Goodman-Bacon A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225, 254–277. DOI: 10.1016/j.jeconom.2021.03.014.

Heckman J., Ichimura H., Todd P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65 (2), 261–294.

Jakiela P. (2021), Simple diagnostics for two-way fixed effects. *ArXiv*:2103.13229. DOI: 10.48550/arXiv.2103.13229.

Kahn-Lang A., Lang K. (2020). The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business & Economic Statistics*, 38 (3), 613–620. DOI: 10.1080/07350015.2018.1546591.

Keele L. J., Small D. S., Hsu J. Y., Fogarty C. B. (2019). Patterns of effects and sensitivity analysis for differences-in-differences. *ArXiv*:1901.01869. DOI: 10.48550/arXiv.1901.01869.

Manski C. F., Pepper J. V. (2018). How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions. *The Review of Economics and Statistics*, 100 (2), 232–244. DOI: 10.1162/REST_a_00689.

Meyer B., Viscusi W. (1995). Workers' compensation and injury duration: Evidence from a natural experiment. *The American Economic Review*, 85 (3), 322–340.

Rambachan A., Roth J. (2023). A more credible approach to parallel trends. *The Review of Economic Studies*, 90, 2555–2591. DOI:10.1093/restud/rdad018.

Rios-Avila F. (2020). Recentered influence functions (RIFs) in Stata: RIF regression and RIF decomposition. *Stata Journal*, 20 (1), 51–94. DOI: 10.1177/1536867X20909690.

Roth J. (2022). Pre-test with caution: Event-study estimates after testing for parallel trends. *The American Economic Review: Insights*, 4 (3), 305–322. DOI: 10.1257/aeri.20210236.

Roth J., Sant'Anna P. H. (2021). Efficient estimation for staggered rollout designs. *ArXiv:2102.01291*. DOI: 10.48550/arXiv.2102.01291.

Roth J., Sant'Anna P. H., Bilinski A., Poe J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235 (2), 2218–2244. DOI: 10.1016/j.jeconom.2023.03.008.

Roth J., Sant'Anna P. H. (2023). When is parallel trends sensitive to functional form? *Econometrica*, 91 (2), 737–747. DOI: 10.3982/ECTA19402.

Rubin D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66 (5), 670–688.

Sun L., Abraham S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225, 175–199. DOI: 10.1016/j.jeconom.2020.09.006.

Villa J. M. (2016). diff: Simplifying the estimation of difference-in-differences treatment effects. *Stata Journal*, 16, 52–71.

Wald A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*, 11, 284–300.

Ye T., Keele L., Hasegawa R., Small D. S. (2021). A negative correlation strategy for bracketing in difference-in-differences. *ArXiv*: 2006.02423. DOI: 10.48550/arXiv.2006.02423.

Received 14.03.2023; accepted 02.10.2023