
СТАТИСТИЧЕСКИЕ МЕТОДЫ И АНАЛИЗ ДАННЫХ



DOI: 10.19181/4m.2025.34.2.1

EDN: VCXNGL

ЦЕЛЕВЫЕ КАУЗАЛЬНЫЕ ЭФФЕКТЫ В СОЦИАЛЬНЫХ ИССЛЕДОВАНИЯХ

Соколов Борис Олегович

Лаборатория сравнительных социальных исследований
им. Р.Ф. Инглхарт Национального исследовательского университета
«Высшая школа экономики»; Петербургский кампус Национального
исследовательского университета «Высшая школа экономики»,
Санкт-Петербург, Россия
Эл. почта: bssokolov@hse.ru
ORCID: 0000-0002-5151-8147

Для цитирования: *Соколов Б. О.* Целевые каузальные эффекты в социальных исследованиях // Социология: методология, методы, математическое моделирование (Социология: 4М). 2025. № 61. С. 7-76. DOI: 10.19181/4m.2025.34.2.1. EDN: VCXNGL.

Данная работа представляет собой обзор основных целевых эффектов (эстимандов), используемых в современных научных и прикладных исследованиях для операционализации каузальных запросов в терминах модели Рубина. Сначала рассматриваются наиболее популярные средние эффекты воздействия (АТЕ, АТТ, АТС). Затем описываются их расширения, в том числе разнообразные локальные и условные эффекты воздействия; целевые величины, квантифицирующие каузальные взаимодействия и каузальную медиацию; эффекты воздействия многозначных и интервальных переменных; а также динамические причинно-следственные связи. Для каждого

из этих эстимандов даётся содержательное объяснение и приводятся примеры практических вопросов, на которые они позволяют ответить. Также обсуждаются допущения, необходимые для непараметрической идентификации ряда эффектов, и методы, которые можно использовать для их оценки.

Ключевые слова: статистический причинно-следственный вывод; эффект воздействия; каузальная модель Рубина; доказательная политика; эстиманд; количественные методы.

Благодарности: Статья подготовлена в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики». Автор выражает признательность за ценные советы и комментарии редактору и анонимному рецензенту, а также Алексею Бессуднову, Егору Лазареву, Елизавете Попковой, Максиму Рудневу и Юрию Рыкову.

1. Введение

Каждый человек, которому хотя бы раз в жизни приходилось принимать решение, связанное с изменением сложившегося порядка вещей (в государственной политике, бизнесе, клинической практике или частной жизни), наверняка задавался вопросами, которые можно свести к базовой условной структуре «*что..., если*». Увеличится ли рождаемость в стране, если перенести основную часть материнского капитала с первого ребёнка на второго? Улучшит ли новый метод лечения выживаемость среди пациентов, заразившихся смертельно опасной инфекцией? Снизится ли смертность или насильственная преступность в регионе, если запретить продавать алкоголь по ночам, и если да, то насколько? Повысит ли продажи ребрендинг товара? Насколько продлит жизнь отказ от курения?

Все эти вопросы объединяют три характеристики. Во-первых, в их фокусе находится причинно-следственный (каузальный¹) эффект некоего (воз-)действия (англ. *treatment*): изменится ли мир нужным нам образом в результате этого действия? Во-вторых, они подразумевают сравнение двух (или более) сценариев (сохранение статус-кво или его изменение в том или ином варианте), которые *ex ante* являются *контрфактуальными* (англ. *counterfactual*), т. е. недоступными прямому наблюдению и измерению. В-третьих, на каждый из них в теории можно дать не только бинарный (*да/нет*), но и количественный ответ, отражающий разницу значений целевой величины между сценариями: 123 516 дополнительных рождений и 3.4% спасенных жизней; 2 864 не случившиеся смерти; 532 миллиона рублей; 9 ½ лет и т.д.

Важно, что именно количественный компонент, если он известен, пусть даже примерно или в качестве априорного предположения, является основополагающим для принятия решения о том, воплощать в жизнь соответствующую «политику»² или нет. Чтобы решить, стоит ли игра свеч, необходимо знать, сколько мы выиграем или проиграем, если выберем конкретную альтернативу. Проблема, однако, заключается в том, что нам не дано в точности предвидеть будущее, а выбрать обычно можно только один вариант и только один раз. Поэтому приходится полагаться

¹ Хотя данный термин, являющийся «калькой» соответствующего английского, не вполне устоялся в отечественной академической традиции, его несомненным преимуществом по сравнению с русским аналогом является краткость. Поэтому я далее использую в основном его (вместе с производными).

² Так как человек — это, как известно, *zoon politikon*, а многие примеры в этом тексте взяты из области государственного управления, далее я использую термин «политика» (в смысле англ. *policy*) в качестве синонима более общего термина «воздействие» (англ. *treatment*), в том числе и применительно к неполитическим сферам общественной жизни.

на неизбежно шумные оценки, препарирующие и суммирующие доступную информацию.

Впрочем, за последние десятилетия экспоненциальный рост объема доступных количественных данных практически во всех сферах человеческой деятельности и постоянное усовершенствование методов их статистической обработки дали учёным — а вслед за ними и прикладным аналитикам — возможность получать всё более точные и надёжные ответы на каузальные вопросы, подобные приведённым выше и представляющие значительный общественный или коммерческий интерес. Хорошей иллюстрацией в этом отношении являются премии Шведского государственного банка по экономическим наукам памяти Альфреда Нобеля 2019 и 2021 годов³, врученные А. Банерджи, Э. Дюфло и М. Кремеру за «экспериментальный подход к борьбе с глобальной бедностью» (фактически — за развитие методологии полевых экспериментов и важные практические результаты в указанной сфере [1; 2]) и Д. Карду, Дж. Энгристу и Г. Имбенсу за «за эмпирический вклад в экономику труда и методологический вклад в анализ причинно-следственных связей» соответственно [4].

Так или иначе, курсы по методам статистического⁴ причинно-следственного вывода (англ. *causal inference*) сегодня читаются во всех ведущих университетах мира; учебники множатся

³ Ряд более ранних лауреатов также внесли значимый вклад в эту область, например, Я. Тинберген и Р. Фриш (1969), Т. Ховельмо (1989), К. Грейнджер (2003) и особенно Дж. Хекман (2000). В работах последнего можно найти подробную историческую справку (напр., [3]).

⁴ Английский оригинал не включает этого уточнения, однако мне представляется принципиально важным обозначить, что речь идёт именно о *статистической*, или *вероятностной* [5], а не *детерминистской* каузальности (которая в большей степени соответствует обыденному пониманию причинно-следственных взаимосвязей).

как грибы, а сами эти методы все более активно используются в разнообразных научных дисциплинах, государственном управлении и бизнесе [6]. Вместе с тем поразительно мало внимания уделяется «наведению мостов» между научными и прикладными задачами, имеющими каузальное измерение, и статистическим инструментарием, используемым для их анализа. В первую очередь, речь идет о такой базовой вещи, как постановка и операционализация исследовательского вопроса, т. е. выражение его в виде некоторой математической *целевой величины* (также называемой *эффектом*, *параметром* или *эстимандом*⁵), которую можно оценить на основе имеющихся данных и которая при этом имеет конкретный содержательный смысл, понятный людям, в премудростях теории вероятностей не сведущим, но облаченным правом принятия решений⁶.

Для иллюстрации этого провокационного тезиса позволю себе обратиться к личной преподавательской практике. В рамках курса по каузальному выводу, который я несколько лет читал в Европейском университете в Санкт-Петербурге, в качестве одного из первых домашних заданий студентам предлагался разбор статей, соприкасающихся с их исследовательскими интересами, использующих обсуждаемые методы и опубликованных в ведущих международных журналах по политологии или социологии. Важным элементом задания было установление целевого эффекта, оцениваемого авторами исходной работы (ATE⁷, SATT,

⁵ Оригинальные английские термины: (*target*) *quantity of interest*, *effect*, *parameter* и *estimand* соответственно.

⁶ Грубо говоря, эта величина должна сообщать примерно следующее: если вы сделаете так, то спасёте (в плохом варианте — убьёте) *n* людей, заработаете *m* денег и т. д. по сравнению со сценарием, в котором вы так не сделаете.

⁷ Смысл этой и следующих двух аббревиатур раскрывается далее по тексту, однако если они уважаемому читателю не очень понятны, то, вероятно, предложенный обзор является актуальным.

LATE и т. д.), соответствующей ему генеральной совокупности (популяции), а также его запись с помощью математической нотации. При этом постоянно возникала проблема: эффект, как правило, сами авторы явно не определяли и тем более не представляли как формулу. Более того, зачастую его было невозможно однозначно вывести из соответствующих фрагментов текста, равно как и сопоставить с коэффициентами, полученными в ходе статистического анализа.

Показательный пример: статья, недавно опубликованная в *American Political Science Review*, одном из наиболее влиятельных журналов по социальным наукам, и имеющая своей целью демонстрацию преодоления методологических трудностей, возникающих при оценке электоральных эффектов случаев массовых расстрелов в американских школах [7], ни разу эти эффекты формально не задаёт. Более того, хотя авторы фокусируются на конкретном статистическом дизайне, методе *разность разностей* (см. ниже, раздел 4.3), они всего лишь раз упоминают в тексте эффект, который является целевым в этом дизайне, *средний эффект воздействия для группы воздействия* (см. определение ниже), и то — цитируя другую публикацию [7, р. 252]. На рисунке 7 в качестве подписи на графике они даже используют обозначение «*средний эффект воздействия*», хотя с точки зрения содержательного смысла и практической применимости это другая величина [7, р. 251]. Таким образом, после прочтения статьи остаются существенные вопросы: эффекты чего, по сравнению с чем, по отношению к чему и для какой совокупности наблюдений пытаются оценивать авторы, и соответствуют ли их результаты исходным интенциям?

Между тем интерпретация и возможность генерализации любой количественной оценки в первую очередь зависят от того, позволяет ли она ответить на вопрос, который интересует исследователя. Если вопрос даже не задан, о чем можно говорить?! Любой статистический метод, сколь угодно продвину-

тый, является валидным только относительно однозначно сформулированной величины, которую он призван оценить, и правдоподобности допущений, на которые он полагается (причём допущения сами во многом завязаны на спецификации целевой величины). Поэтому обязательным условием успешности любого исследования (неважно — научного или прикладного), ставящего каузальный вопрос, является демонстрация тождественности вопроса, высказанного на естественном языке, и его «перевода» на язык статистической модели (т. е. целевой величины/эффекта). Без подобной демонстрации невозможно даже понять, способны ли вообще данные ответить на исходный вопрос, не говоря уже о том, что в случае несоответствия между содержательным вопросом и вопросом, на который фактически отвечает полученная оценка, любые последующие практические решения представляются в лучшем случае необоснованными, а в худшем — могут нанести вред.

Процитированная статья — далеко не единственный пример пренебрежения этим принципом, вроде бы простым и понятным до самоочевидности, в современных социальных науках. Возможно, проблема в том, что в подавляющем большинстве учебников, онлайн-пособий и методологических статей, посвященных каузальному выводу, в фокусе находятся допущения, вычислительные методы и детали их имплементации в том или ином программном пакете, а целевой эффект декларируется в самом начале и далее полагается, что его прикладной смысл ясен для читателя. На мой взгляд, это сомнительное предположение.

Пожалуй, единственная известная мне публикация, в которой авторы развернуто объясняют основные целевые эффекты, имеющие каузальную интерпретацию, — это до сих пор не вышедший в рецензируемом журнале препринт Н. Грейфера и Э. Стюарт [8], который к тому же содержательно привязан к одному конкретному семейству аналитических техник (мэтчинг и

обратное взвешивание по вероятности)⁸. Даже в статье Лундберга и соавторов [11], которая своим названием (*What Is Your Estimand?*) подчеркивает важность четкого определения целевого эффекта и предлагает подробную инструкцию, не содержится подробного описания сути упоминаемых эстимандов и их соотношения с содержательными исследовательскими вопросами.

Настоящая работа призвана восполнить обозначенный пробел. В ней предлагается обзор основных целевых эффектов, которые используются в современных научных исследованиях и бизнес-аналитике для операционализации каузальных запросов с помощью так называемой каузальной модели Рубина. Для каждого из рассматриваемых эффектов даётся содержательное объяснение и приводятся примеры практических вопросов, на которые они позволяют ответить.

Статья организована следующим образом. В следующем разделе вводятся основные идеи и термины указанной модели. Затем описываются наиболее часто используемые средние эффекты воздействия (АТЕ, АТТ, АТС) и обсуждаются допущения, требуемые для их успешной эмпирической идентификации. После этого рассматриваются расширения базовых эффектов, в том числе разнообразные локальные и условные эффекты воздействия; целевые величины, позволяющие квантифицировать каузальные взаимодействия и каузальную медиацию; эффекты воздействий многозначных и интервальных переменных; а также динамические причинно-следственные связи.

⁸ Стоит также указать на ряд онлайн-источников, которые не являются официальными академическими публикациями, но предлагают довольно подробное и информативное введение в тему каузальных эстимандов. Например, заметка в блоге Э. Хайса [9] или глава 10 в электронном учебном пособии М. Барретта и коллег [10].

2. Начала

2.1 Каузальная модель Рубина

Данная модель⁹ [19; 20; 21; 22; 23] является наиболее влиятельной¹⁰ в современной статистике парадигмой для работы

⁹ Американский статистик Дональд Рубин не является «первооткрывателем» основных идей модели, получившей его имя. Например, центральное для неё понятие потенциальных исходов было, по всей видимости, введено Ежи (Юрием) Нейманом [12], а его философские корни можно проследить как минимум до Юма и Лейбница. Однако именно в работах Рубина контрфактуальный подход к статистическому причинно-следственному выводу получил относительно законченный вид. Кроме того, Рубин распространил его на обсервационные (англ. *observational*) исследования: до этого потенциальные исходы в эксплицитном виде использовались только в методологической литературе по рандомизированным экспериментам (хотя схожие идеи высказывались также в психометрике и эконометрике). Сам термин «модель Рубина» предложен П. Холландом [13]. Подробнее об истории вопроса можно узнать в следующих источниках [3; 14; 15; 16; 17; 18].

¹⁰ Но не единственной. Её основной альтернативой, в последние годы набирающей всё большую популярность, является теория непараметрических структурных уравнений (англ. *non-parametric structural equation models*), которая также известна как теория структурных каузальных моделей (англ. *structural causal models*) и теория графических каузальных моделей, или НАГов (*направленных ациклических графов*; от англ. *directed acyclic graphs* — DAGs), и в наибольшей степени ассоциируется с именем и работами израильско-американского статистика и информатика Джуды Пёрла [24; 25].

Впрочем, как неоднократно отмечал сам Пёрл, аксиоматизация каузальных проблем в рамках этой теории логически эквивалентна подходу Рубина: «два исследователя, работающие с одним и тем же вопросом на тех же данных и с теми же допущениями, получают одинаковый результат независимо от того, используют ли они модель Рубина или модель, основанную на структурных уравнениях. Теорема в одной [из этих моделей; курсив мой — прим. авт.] является теоремой в другой. Допущение в одной является допущением в другой» [18, р. 30; 26, р. 314] (с другой стороны, Пёрл может переоценивать близость этих парадигм [27]).

Кроме того, модель Рубина гораздо чаще встречается на практике в большинстве социальных наук [28], за исключением разве что эпидемиологии, да и в

с каузальными исследовательскими вопросами. Существует значительное число учебников, предлагающих различающиеся по сложности и глубине введения в неё. Некоторые из них находятся в открытом доступе и не требуют от читателя полноценного математического образования, поэтому вполне могут быть рекомендованы тем, кто только начинает знакомиться с указанной проблематикой [31;32; 33; впрочем, это далеко не полный список]. На русском языке описание этой модели можно найти в ряде обзорных статей [2; 34; 35; 36] и учебных пособий [37].

Ключевым понятием модели Рубина являются потенциальные исходы (англ. *potential outcomes*), то есть гипотетические значения зависимой переменной Y для отдельных наблюдений i , соответствующие различным возможным значениям воздействующей переменной (англ. *treatment*) D . Для простоты рассмотрим базовый сценарий, в котором последняя может принимать только два значения (самый известный вариант такого сценария — это классический рандомизированный эксперимент): 1, если воздействие имеет место (так называемая *группа воздействия*¹¹ — англ. *treatment group*), и 0, если нет (*группа контроля/контрольная группа* — англ. *control group*).

последней инструменты структурного подхода, как правило, используются не для операционализации исследовательских запросов, а для визуальной иллюстрации аналитических допущений. Наконец, контрфактуальная нотация применяется для задания целевых эффектов и в теоретических работах [29], и даже в учебных пособиях по структурным каузальным моделям [30, глава 4]. Поэтому я предпочитаю использовать более популярную и влиятельную в целом (равно как и более знакомую отечественному читателю) традицию. Тем не менее многие (хотя не все) из эстимандов, обсуждаемых в настоящей работе, могут быть заданы и в терминах *do*-исчисления (англ. *do-calculus*) Пёрла.

¹¹ Седашов [36] переводит *treatment group* как «экспериментальная группа». На мой взгляд, такой перевод является не вполне удачным: зачастую для поисков ответа на каузальные вопросы используются обсервационные, а не экс-

Например, мы хотим понять, повышает ли вероятность оформления платной подписки на некоторый онлайн-сервис наличие рисунка милого котика на всплывающем сообщении с соответствующим предложением. Предположим, что пользователь i любит котов, так что для него такой рисунок станет весомым эмоциональным аргументом в пользу оформления подписки, а вот без гипнотического воздействия пушистой зверушки на его подсознание он не готов платить за услуги сервиса. В таком случае потенциальные исходы *под воздействием*¹² ($Y_i^{D=1}$) и *без воздействия* ($Y_i^{D=0}$)¹³ для i будут равняться 1 (кот есть → подписка оформлена) и 0 (нет кота → подписка не оформлена). Пользовательница i' , в свою очередь, котов не любит, равно как и лишними деньгами не обладает, поэтому для неё оба потенциальных исхода равняются нулю (подписка не оформляется в любом случае).

В этом примере легко вычислить, что для i индивидуальный эффект воздействия (англ. *individual treatment effect* — ITE), то есть предъявление кота, является положительным, тогда как для i' — нулевым. Можно представить и других пользователей, для которых этот эффект будет отрицательным. Если мы знаем

периментальные данные. Странно называть что-то «экспериментальным», если фактически эксперимент не ставился. Термин «группа воздействия» является более общим и позволяет характеризовать как классические эксперименты, так и обсервационные исследования.

¹² Термин *treatment* на русский можно также перевести как *вмешательство* [36].

¹³ Существуют различные варианты нотации для обозначения потенциальных исходов. Пожалуй, наиболее популярными являются следующие два: $Y^d \{Y^1, Y^0, \dots\}$ или $Y(d) \{Y(1), Y(0), \dots\}$. Я, однако, предпочитаю эксплицитно включать в нотацию указание на то, что 1 и 0 (или любые другие) — это значения воздействующей переменной D , а не чего бы то ни было ещё, поэтому использую вариант $Y^{D=d}$. Так получается более громоздко, но зато снижается риск недопонимания со стороны неискушенных читателей.

потенциальные исходы для каждого пользователя, то можем легко настроить таргетированную рекламу и максимизировать число платных подписчиков, одновременно не раздражая тех, кто негативно относится к котам (или вылечить всех пациентов, которым помогает новое лекарство, и не навредить тем, для кого оно чревато опасными побочными эффектами; или оптимально распределить государственные средства на реализацию программы помощи бедным в те регионы, где эта программа будет иметь эффект; и т.д.). Проблема в том, что мы практически никогда не можем наблюдать все возможные потенциальные исходы. Обычно мы видим только один (соответствующий наблюдаемому значению переменной D), а второй (или все остальные, если переменная D является многозначной) является пропущенным значением. Данное затруднение, известное как *фундаментальная проблема каузального вывода* [13, р. 947; 38, р. 7], делает оценку индивидуальных эффектов воздействия практически неразрешимой задачей (впрочем, ниже мы ещё вернемся к этому вопросу).

С другой стороны, зачастую для принятия решений вовсе не обязательно знать, как изменение статус-кво повлияет на каждого затронутого им индивида или каждый затронутый объект. Информация о том, насколько больше пациентов выживет, если текущую терапию острой инфекции (например, лихорадки Эбола) заменить на новую, насколько увеличится число платных подписчиков онлайн-сервиса, если им показывать кота, или насколько уменьшится смертность в регионе, если запретить продажу алкоголя после 21 часа (вместо 22 часов), представляет вполне конкретный практический интерес даже в усреднённом виде. Поэтому в качестве *целевых величин/эстимандов* для прикладных исследователей выступают разнообразные агрегированные эффекты воздействия, усредняющие ITE по тому или иному сегменту изучаемой генеральной совокупности (ГС).

2.2 Основные средние эффекты воздействия

Самый популярный (и самый базовый) вариант каузальной целевой величины, позволяющий «оцифровать» вопросы наподобие тех, что приводятся в предыдущем абзаце, так и называется — *средний (усреднённый) эффект воздействия* (англ. *average treatment effect* — АТЕ). Он представляет собой математическое ожидание распределения индивидуальных эффектов воздействия по всем представителям совокупности: $\tau_{ATE} = E(Y_i^{D=1} - Y_i^{D=0})$. Содержательно эту величину можно интерпретировать следующим образом: насколько в среднем по популяции изменится значение целевой метрики при имплементации соответствующего воздействия в качестве универсальной политики.

Наиболее частой альтернативой АТЕ выступает АТТ, *средний эффект воздействия в группе, подвергнувшейся воздействию* (или просто *в группе воздействия*; англ. *average treatment effect on the treated*). Главное отличие между этими двумя величинами состоит в том, что АТЕ усредняет по всем наблюдениям в ГС, тогда как АТТ — только по тем, которые на самом деле испытали воздействие: $\tau_{ATT} = E(Y_i^{D=1} - Y_i^{D=0} | D_i = 1)$. Другими словами, АТТ показывает, как воздействие в среднем повлияло на тех, на кого оно было оказано, по сравнению с контрфактуальной ситуацией, в которой воздействия не было [8, р. 6].

Помогла ли новая методика лечения тем, на ком она испытывалась? Повлияла ли программа повышения квалификации на продуктивность труда *тех сотрудников, которые в ней участвовали* [39]? Способствуют ли ограничения на продажу алкоголя снижению насильственной преступности или росту ожидаемой продолжительности жизни *в тех регионах, в которых они вводятся*? На эти содержательные вопросы можно дать количественный ответ, используя АТТ в качестве целевой метрики. Необходимо ещё раз подчеркнуть, что эти вопросы касаются

в основном *эффективности уже совершенного воздействия по отношению к тем, на кого оно было оказано*. Если полученная оценка АТТ показывает, что воздействие имело отрицательный или нулевой эффект, то целесообразность его продолжения можно поставить под сомнение.

Однако если воздействие оказалось положительным, это не означает, что его непременно нужно распространять на остальных представителей популяции: имеющиеся в распоряжении исследователя свидетельства ничего не говорят об эффекте воздействия для тех, кто ему не подвергся. В этом отношении АТЕ является более общей величиной, так как показывает ожидаемое изменение целевой метрики при имплементации воздействия для всех наблюдений в ГС независимо от того, подвергались ли они реальному воздействию в ходе проведения исследования/пилота. В силу этого АТЕ позволяет отвечать на более общие содержательные вопросы, которые к тому же позитивно окрашены: *следует ли распространить некоторую политику на всю популяцию (а не следует ли прекратить некоторую политику, которая касается отдельных представителей популяции — как в случае с АТТ)*.

Ещё одной родственной АТЕ и АТТ величиной является средний эффект воздействия в контрольной группе — АТС (англ. *average treatment effect on the control* ¹⁴): $\tau_{ATC} = E(Y_i^{D=1} - Y_i^{D=0} | D_i = 0)$. По аналогии с АТТ этот показатель отражает, насколько (не)благоприятным оказалось отсутствие воздействия для контрольных наблюдений. В целом все содержательные вопросы, на которые можно ответить с помощью данной величины, сводятся к формуле *«следует ли распростра-*

¹⁴ В некоторых источниках (в частности, у Грейфера и Стюарт [8]) вместо *control* используется *untreated* — «не подвергшиеся воздействию». Аббревиатура меняется соответственно — АТУ вместо АТС.

нить некоторую политику на тех, кого она до сих пор не касалась» [8, р. 6].

При определенных условиях все три целевых величины могут быть эквивалентными друг другу. Например, в истории имели место случаи выживания после падения с большой высоты без парашюта (равно как и гибели при его наличии). Однако вряд ли кто-то сможет убедительно оспорить утверждение о том, что разница вероятностей остаться в живых при наличии парашюта и при его отсутствии (т. е. АТЕ) близка к единице, причём не только в среднем, но и для каждого отдельно взятого индивида¹⁵. Более того, в этом примере АТТ (на сколько уменьшится вероятность выживания прыгающих, если у них предварительно отберут парашют) равен АТС (на сколько увеличится вероятность выжить у падающих с высоты, если им в полёте некие высшие силы любезно выдадут парашют), и обе эти величины также равны АТЕ. Однако на практике индивидуальные эффекты воздействия являются гораздо более вариабельными (одно и то же лекарство кому-то помогает избавиться от головной боли, а кому-то нет), в силу чего АТЕ, АТТ и АТС могут различаться, причем зачастую — достаточно сильно.

Впрочем, в рандомизированных экспериментах эта проблема стоит не так остро: если сама выборка является случайной и достаточно большой, то благодаря рандомизации группы воздействия и контроля также являются случайными выборками из целевой совокупности. Таким образом, средние по индивидуальным эффектам воздействия в каждой группе (т. е. АТТ и АТС) будут примерно равны АТЕ даже при наличии индивидуальной гетерогенности. Однако во многих сферах рандомизиро-

¹⁵ Этот пример вдохновлён известной заметкой Г. Смита и Дж. Пелл [40; см. также 41].

ванные эксперименты организовать тяжело: препятствуют разнообразные логистические, финансовые, этические и иные ограничения. Кроме того, рандомизация невозможна и при анализе уже собранных ранее неэкспериментальных, или обсервационных (англ. *observational*), данных: накопленной экономической, демографической или медицинской статистики, пользовательских данных онлайн-магазина или социальной сети и т. д.

Отсутствие рандомизации накладывает существенные ограничения на спектр вопросов, ответы на которые можно найти с помощью статистического анализа: получить надёжные количественные оценки АТЕ во многих таких контекстах попросту не получится. Поэтому в качестве целевых эстимандов приходится выбирать АТТ, АТС или разнообразные локальные эффекты воздействия (о которых речь пойдёт в следующем разделе), жертвуя внешней валидностью ради большей надёжности.

2.3 Проблема идентификации и необходимые допущения

Когда и почему в обсервационных исследованиях используются АТТ или АТС вместо АТЕ? Здесь нужно обратиться к основам статистического причинно-следственного вывода. Чтобы получить точную количественную оценку АТЕ (как, впрочем, и любого другого каузального эстиманда, который упоминается в этом тексте), исследователю необходимо удостовериться самому и убедить других в том, что на имеющихся данных можно уравнивать следующие три величины: (1) собственно целевой эффект воздействия τ ; (2) его эмпирический эквивалент θ ; (3) и выборочную оценку последнего $\hat{\theta}$ [11; ср. 3, р. 2].

Целевой, или теоретический, эстиманд τ всегда задаётся в контрфактуальных (то есть ненаблюдаемых и потому не поддающихся измерению) терминах. Для каждой единицы анализа известен лишь реализовавшийся потенциальный исход, а о значениях остальных можно только гадать. Для вычисления τ тре-

буется доступ как к каждому элементу генеральной совокупности (что в теории вероятно для конечных ГС небольшого размера), так и к возможному миру (англ. *possible world*), в котором D для данного элемента имеет не то значение, которое фиксируется в действительности (что пока вообразимо только в фантастических произведениях). В лучшем случае можно попробовать аппроксимировать распределения потенциальных исходов, для чего и требуется перевод в теоретически доступные наблюдению величины и вынос контрфактуальных терминов за скобки, или *идентификация* (англ. *identification*).

θ (Лундберг и соавторы [11] называют эту величину *эмпирическим эстимандом*) определяется в терминах потенциально наблюдаемых величин, характеризующих целевую популяцию. Даже если представить рандомизированный эксперимент, который охватил абсолютно всех жителей России, мы все равно не можем измерить $Y_i^{D=0}$ для тех, кто попал в группу воздействия, и $Y_i^{D=1}$ для тех, кто попал в группу контроля. Однако мы можем посчитать $E(Y_i|D_i = 1)$ и $E(Y_i|D_i = 0)$. Так как воздействие назначалось случайным образом, то $E(Y_i|D_i = 1) = E(Y_i^{D=1})$ и $E(Y_i|D_i = 0) = E(Y_i^{D=0})$. Проще говоря, вычислив средние значения Y в группах контроля и воздействия, мы получим несмещённые оценки средних значений потенциальных исходов под воздействием и без него.

Эта же логика работает и на втором этапе (установление тождества θ и $\hat{\theta}$): если мы имеем случайную выборку из ГС, то выборочные средние в группах воздействия и контроля являются несмещёнными и состоятельными оценками соответствующих популяционных параметров, а значит — и исходных контрфактуальных средних, необходимых для вычисления АТЕ.

Необходимо отметить, что контролируемая исследователем рандомизация сама по себе не является необходимым

условием успешной идентификации¹⁶ целевого эффекта (то есть возможности его перевода из контрфактуальных выражений в потенциально наблюдаемые популяционные параметры). В целом знак равенства между ожиданиями потенциальных исходов и популяционными средними в группах воздействия и контроля можно поставить тогда, когда выполняются допущения, обычно обозначаемых как SUTVA¹⁷, плюс следующие *идентифицирующие* условия: *взаимозаменяемость* (англ. *exchangeability*¹⁸) и *пересечение/позитивность* (англ. *overlap/positivity*) [46].

¹⁶ С другой стороны, важно подчеркнуть, что успешная идентификация целевого эффекта не означает с необходимостью возможность его успешной *оценки* на конечной выборке. Например, если находящаяся в распоряжении исследователя выборка является смещенной относительно ГС по целевым параметрам (или по характеристикам, которые могут выступать в качестве общих причин D и Y) или страдает от ошибок измерения (например, социальной желательности), то наблюдаемая разница средних между группами воздействия и контроля будет отличаться от таковой в ГС и, соответственно, будет давать смещённую оценку как θ , так и τ .

¹⁷ SUTVA — от англ. *stable unit-treatment value assumption*. Это набор априорных допущений, позволяющих однозначно соотнести потенциальные и наблюдаемые исходы. Существуют разные формулировки SUTVA. Практически во всех из них упоминаются следующие основные пункты: (1) согласованность (англ. *consistency* — не путать с *состоятельностью* оценки, которая по-английски обычно обозначается тем же словом) — наблюдаемый исход $Y_i(D_i = d)$ тождественен потенциальному исходу $Y_i^{D=d}$ для соответствующего значения D ; (2) эквивалентность воздействий (всем дают лекарство с идентичным химическим составом; всех пациентов оперируют врачи с одинаковыми навыками и в одинаковых условиях); (3) отсутствие экстерналий/побочных эффектов (англ. *spillover effects*): воздействие на наблюдение i не связано с потенциальными исходами для других наблюдений i' [13; 42; 43; 44; 45].

¹⁸ Данное условие имеет множество вариантов и названий; основными альтернативами в англоязычной традиции являются следующие: *ignorability*, *exogeneity* и *unconfoundedness*. Я предпочитаю термин *exchangeability* и рус-

Взаимозаменяемость требует, чтобы потенциальные исходы и назначение воздействия не были связаны с друг другом: $(Y_i^{D=1}, Y_i^{D=0}) \perp\!\!\!\perp D_i$ (символ $\perp\!\!\!\perp$ обозначает статистическую независимость). Содержательный смысл таков: зная значения потенциальных исходов для i -го наблюдения, мы тем не менее не можем предсказать значение D_i (и наоборот). Таким образом, у потенциальных исходов и воздействующей переменной отсутствуют общие причины, в силу чего любую разницу уже в наблюдаемых исходах (которые, согласно допущению о последовательности, входящему в SUTVA, представляют собой реализовавшиеся потенциальные исходы для соответствующего значения D — см. примечание 19) между группами воздействия и контроля мы можем атрибутировать самому факту воздействия.

В абстрактном сферическом эксперименте в вакууме это очевидно: единственной причиной, порождающей наблюдаемое распределение D , является алгоритм, лежащий в основе используемого исследователем генератора квазислучайных чисел. Этот алгоритм вряд ли имеет какое-то отношение к распределению потенциальных исходов, что позволяет отсечь основные альтернативные объяснения наблюдаемой статистической связи между Y и D — обратную каузальность (Y влияет на D , а не наоборот) и *конфаундинг* (англ. *confounding*), или наличие общих причин, создающих ложную корреляцию между целевыми признаками.

ский перевод «взаимозаменяемость», потому что он, на мой взгляд, наилучшим образом передаёт ключевую интуицию, лежащую в основе этого идентифицирующего допущения: если назначение воздействия не зависит от потенциальных исходов, то группы воздействия и контроля являются *взаимозаменяемыми* в том смысле, что мы можем использовать информацию о наблюдаемом (реализованном) распределении Y в одной из них для восстановления ненаблюдаемой информации о распределении потенциальных исходов для соответствующего условия $Y^{D=d}$ в другой.

В обсервационных исследованиях, как правило, делается допущение об условной (по ковариатам) взаимозаменяемости: $(Y_i^{D=0}, Y_i^{D=1}) \perp\!\!\!\perp D_i \mid X_i$. Другими словами, мы предполагаем¹⁹, что содержательный механизм назначения воздействия нам известен, и за счёт эксплицитного включения в статистическую модель факторов, в этом процессе задействованных (т. е. конфаундеров X), мы «обнуляем» влияние общих причин Y и D , так

¹⁹ Необходимо отметить, что это очень сильное допущение. На практике мы практически никогда не можем учесть все возможные общие причины D и Y , либо потому что не можем их измерить в силу тех или иных ограничений, либо потому что даже не догадываемся об их существовании. С другой стороны, в хорошо изученных контекстах основные конфаундеры, как правило, известны, так что их включение в используемую для оценки целевого эффекта статистическую модель не представляет особых сложностей. При этом искажения оценки искомого параметра, возникающие из-за влияния оставшихся за рамками модели факторов, не всегда являются критичными в том смысле, что они как минимум не ведут к изменению знака или потере статистической значимости полученной оценки, а иногда даже её абсолютной величины.

Кроме того, с помощью семейства вспомогательных статистических методик, известных под общим названием «анализ чувствительности» (англ. *sensitivity analysis*), можно вычислить, насколько сильным должно быть влияние неучтенных конфаундеров, чтобы изменить содержательные выводы относительно величины и направления эффекта, и затем (на основе накопленных результатов предыдущих исследований, теории и/или здравого смысла) решить, насколько вероятно наличие подобной пропущенной переменной в ситуации конкретного обсервационного исследования [47; 48; 49].

Так, в середине прошлого века неоднократно выдвигался (в том числе такими признанными авторитетами, как Р. Фишер [50]), аргумент о том, что наблюдаемая корреляция между курением и развитием рака лёгких объясняется влиянием неизвестных общих причин, например генетической предрасположенности. Однако затем было показано, что для полного «обнуления» имеющихся оценок этой корреляции гипотетический «ген курения» должен повышать риск заболевания примерно в девять (9) раз и при этом встречаться среди курильщиков в те же девять раз чаще [51]. На тот момент (как и сейчас) науке такие гены известны не были, и это послужило одним из окончательных доказательств каузального характера связи вредной привычки и рака.

что D всё ещё остается единственным релевантным показателем, объясняющим разницу в значениях Y между группами контроля и воздействия.

Допущение о пересечении²⁰, в свою очередь, требует ненулевой (равно как и не единичной) исходной вероятности воздействия для всех наблюдений: $0 < P(D_i = 1) < 1 \forall i$. В своей условной версии — независимо от их значений по ковариатам: $0 < P(D_i = 1 | X_i = x) < 1 \forall i$ и x . Величина $P(D_i = 1 | X_i = x)$ известна как *мера склонности* (англ. *propensity score*) и часто обозначается как $e(X)$. Если $e(X)_i$ равняется 0 или 1, то, во-первых, возникает философский вопрос о том, можно ли соответствующие потенциальные исходы ($Y_i^{D=1}$ в первом случае и $Y_i^{D=0}$ во втором), которые не имеют шансов реализоваться на практике, считать «потенциальными»? Осмысленность ITE для подобных наблюдений также оказывается сомнительной: зачем спрашивать «что ..., если», если вероятность «если» априорно нулевая [52, р. 141–142]? Более того, когда наблюдения с $X_i = x$ попадают только в группу контроля или только в группу воздействия, для них в любом случае нельзя рассчитать подвыборочный, или условный, средний эффект воздействия (см. подробное обсуждение этого типа эстимандов ниже), что снижает репрезентативность и точность оценки основной целевой величины²¹ [53; 54; 55].

Во-вторых, зачастую на практике нарушение допущения о пересечении означает и проблемы с взаимозаменяемостью. В неформальном виде эту интуицию можно представить сле-

²⁰ Другой вариант русского перевода — «перекрытие».

²¹ Если условная взаимозаменяемость и условное пересечение выполняются, АТЕ можно представить как взвешенное среднее стратоспецифичных эффектов $\tau_x = E(Y_i^{D=1} - Y_i^{D=0} | X_i = x)$ по всем возможным профилям (комбинациям уникальных значений переменных, входящих в множество X) ковариат x .

дующим образом: раз наблюдения с определёнными характеристиками чаще оказываются в группе контроля или воздействия, то сравниваемые группы в итоге различаются не только по значению D , но и по значениям X . Таким образом, мы уже не можем воспользоваться вероятностным аналогом принципа единственного различия, описанным ранее: разница в средних между сравниваемыми группами может частично или даже полностью объясняться различиями не по статусу воздействия, а по другим наблюдаемым или ненаблюдаемым характеристикам (так называемым *дисбалансом по ковариатам*²² — англ. *covariate imbalance*).

2.4 ATE, ATT или ATC: как выбрать?

Здесь, наконец, можно вернуться к проблеме выбора целевого эффекта в обсервационных исследованиях. Как отмечено выше, в экспериментах воздействие назначается случайным образом. Поэтому потенциальные исходы автоматически оказываются ортогональными D . Допущение о пересечении также выполняется: вероятность (не)попадания в группу воздействия контролируется исследователем и имеет одно и то же значение для всех участников (как минимум в пределах заданных страт, если речь о стратифицированном эксперименте). Однако в большинстве обсервационных исследований имеет место неслучайный (само)отбор.

Например, искусственная вентиляция легких назначается пациентам с COVID-19 в худшем состоянии²³. Их шансы уме-

²² В общем случае дисбаланс по ковариатам можно определить как наличие любых расхождений в одно- и многомерных распределениях переменных, входящих в X , между группами воздействия и контроля. Например, чаще всего уровень дисбаланса выражается через различия в стандартизированных средних отдельных ковариат, хотя существуют (и во многих случаях являются уместными) и более изощрённые метрики [56].

реть изначально намного выше, чем у тех, кто чувствует себя более-менее нормально и лечится таблетками, прописанными участковым терапевтом. Если доля умерших среди попавших на ИВЛ, превышает таковую среди тех, кого на вентиляцию не отправляли, это не означает, что ИВЛ повышает смертность от коронавируса: просто и вероятность назначения данного вида терапии, и вероятность умереть у таких пациентов исходно были выше из-за наличия общей причины (плохого состояния). Допущение о взаимозаменяемости в этом примере очевидно нарушается: назначение D положительно коррелирует с вероятностью смерти в отсутствие ИВЛ ($Y^{D=0}$). С допущением о пересечении также возникают проблемы: люди с легкой формой заболевания имеют близкую к нулю вероятность попадания на ИВЛ, тогда как с тяжелой — близкую к единице. При этом между первыми и вторыми существуют другие важные различия: возраст, наличие сопутствующих заболеваний и т. д.

Таким образом, для оценки целевого каузального эффекта в неэкспериментальном исследовании необходимо каким-то образом учесть («заблокировать», если пользоваться терминологией Дж. Перла) эти дополнительные различия по ковариатам. Зачастую это делается путём восстановления ненаблюдаемых контрфактуальных исходов за счёт создания синтетической группы сравнения, в которой распределение релевантных ковариат (или хотя бы его основные моменты) совпадает с таковым в целевой группе. Синтетическая группа формируется из числа контрольных наблюдений, если целевой эффект — АТТ, или из числа наблюдений в группе воздействия, если целевой эффект — АТС. Для этого обычно используются такие методы, как мэтчинг, обратное взвешивание по вероятности воздейст-

²³ Этот пример является злободневной адаптацией классической истории про госпитализацию из «Почти безвредной эконометрики» [57].

вия, дважды устойчивая оценка²⁴ и т. д. Общее название данного семейства — отбор по наблюдаемым характеристикам (англ. *selection on observables*).

На практике число наблюдений под воздействием часто существенно меньше числа наблюдений в контрольной группе. При этом синтетическую группу сравнения проще сформировать из более крупной по размеру: в таком случае есть шансы найти более-менее точные пары для отдельных наблюдений (при использовании мэтчинга) или добиться общевыборочного баланса за счёт присвоения меньших весов «экстремальным» контрольным наблюдениям (при использовании взвешивания).

Например, если воздействию (подключению к аппарату ИВЛ) подвергается один пациент из десяти, то для подбора его наиболее точного контрфактуального «близнеца» с помощью той или иной разновидности мэтчинга для последующей оценки АТТ нам доступен пул из девяти наблюдений-кандидатов. Если нас интересует АТС, то в данном примере придётся использовать единственное наблюдение (куривший с дошкольного

²⁴ Во многих из этих методов мера склонности используется для подбора ответов или при расчёте весов индивидуальных наблюдений. В первом случае наблюдениям с нулевыми (близкими к нулю) или единичными (близкими к единице) мерами склонности зачастую невозможно найти «пару», в силу чего их приходится исключать из расчётов. Так как это исключение происходит не случайным образом, а условно по X , итоговый набор данных отличается по своим характеристикам и от исходной выборки, и от целевой генеральной совокупности, а полученная эмпирическая оценка не может претендовать на статус популяционной (см. раздел про FSATT и АТО ниже). Во втором случае мера склонности используется в знаменателях соответствующих формул, в том числе в варианте $1 - e(X)_i$, так что экстремальные значения этой величины транслируются в чрезмерно большие или даже бесконечные индивидуальные веса [32, р. 25; 46, р. 3; 52, р. 141]. Указанные проблемы хорошо иллюстрируют принципиальную роль допущения о пересечении в обсервационных исследованиях, которые полагаются на контроль по наблюдаемым характеристикам.

возраста мужчина 86 лет с раком лёгких, проработавший всю жизнь в угольной шахте) под воздействием как референтную пару для восстановления пропущенного $Y^{D=1}$ для каждого из девяти контрольных наблюдений (среди которых есть профессиональные спортсмены). Насколько надежными будут результаты подобных сопоставлений — вопрос риторический. Впрочем, выбор между АТТ и АТС не обуславливается исключительно соображениями доступности данных: содержательная сторона вопроса, интересующего исследователя или лицо, принимающее решение, при прочих равных имеет приоритет.

3. Локальные, условные и достижимые эффекты воздействия

Зачастую условия, необходимые для идентификации исходного целевого эффекта, не выполняются. В таком случае имеет смысл переопределить эстиманд, задав его для некоторой группы наблюдений, относительно которых можно предложить убедительное эмпирическое и/или содержательное обоснование идентифицирующих допущений. В этом случае мы жертвуем внешней валидностью (оценка целевого эффекта не является репрезентативной для популяции в целом) ради шанса получить относительно точные и надёжные результаты для выбранной страты.

Такие эстиманды называются *локальными*. Популярными методами оценки локальных эффектов воздействия являются метод инструментальных переменных [58] и разрывный дизайн²⁵ (англ. *regression discontinuity design* [59; 60]).

²⁵ Перевод «разрывная регрессия», который встречается в отечественных работах, представляется не вполне точным: в оригинале имеется в виду наличие разрыва в регрессионной функции в точке назначения воздействия. В таком

3.1 LATE в инструментальной оценке

При использовании инструментальных переменных целевой величиной является так называемый средний эффект воздействия среди «послушных» (англ. *compliers*), или CATE. Я, однако, далее буду обозначать его как LATE–IV, чтобы избежать путаницы с более общим семейством эстимандов, которые также часто маркируют аббревиатурой CATE — условными средними эффектами воздействия (англ. *conditional average treatment effects*). Собственно, LATE–IV является частным случаем CATE. В сценарии с бинарной D и бинарным *инструментом* Z LATE–IV можно определить как: $\tau_{\text{LATE-IV}} = E(Y_i^{D=1} - Y_i^{D=0} | U_i = c)$. В этом выражении U_i — фиктивная переменная (по-английски её обычно называют *principal strata*), которая характеризует ненаблюдаемый на практике уровень *послушания* (англ. *compliance*) конкретного наблюдения. Она может принимать четыре значения: *послушный, всегда непослушный, всегда за и всегда против*. Соответствующие английские термины — *complier (c), defier (d), always-taker (a) и never-taker (n)*.

Чтобы понять смысл этих терминов, представим распространённую в клинических испытаниях (и ещё более — в клинической практике) ситуацию несоблюдения предписаний исследователей или лечащих врачей. Например, если суть терапии заключается в том, что надо пить невыносимо горькие таблетки 5 раз в день в течение двух недель, причём участники исследования находятся не в больнице, а дома, и имеют возможность заниматься повседневными делами или работают, то наверняка некоторые из них будут отклоняться от инструкции: потому что лекарство неприятное; потому что оно даёт побочные

варианте корректнее было бы переводить «разорванная регрессия». Дословный перевод «дизайн, основанный на разрыве регрессии», в свою очередь, является слишком громоздким.

эффекты; потому что много работали или просто запомнили; по каким-то другим причинам. Такие участники изначально приписаны к группе воздействия, т. е. для них $Z_i = 1$, однако $D_i = 0$ — по факту терапия не соблюдается и воздействие не оказывается в том виде, в котором оно определено.

Все участники, у которых $D_i \neq Z_i$, автоматически относятся к *непослушным*. Однако указанная категория может включать и тех, у кого $D_i = Z_i$. Всё дело в том, что основанием для классификации является не наблюдаемое значение D , а потенциальное — $D^{Z=z}$. Потенциальные воздействия определяются точно так же, как и потенциальные исходы: для дихотомического инструмента существует два потенциальных воздействия $D^{Z=1}$ и $D^{Z=0}$. Формально *послушными* называются те наблюдения, у которых потенциальные воздействия совпадают с соответствующими значениями инструмента: $D^{Z=1} = 1$ и $D^{Z=0} = 0$.

У непослушных в узком смысле (*defiers*) потенциальное воздействие всегда отличается от значения инструмента: $D^{Z=1} = 0$ и $D^{Z=0} = 1$. Можно сказать, что это те, кто всегда поступает противоположным предписаниям способом. Педагогическая практика показывает, что хорошей (хотя и не вполне точной) иллюстрацией понятия *всегда непослушных* являются сложные отношения котов и дверей. Даже если у вас нет кота, вам наверняка попадались на глаза характерные видео, в которых усатого проказника до неистового мяуканья раздражает закрытая дверь ($Z_i = 0$): он хочет в неё войти во что бы то ни стало и всеми силами привлекает внимание присутствующих к себе и своей потребности. Намерения кота не совпадают с предписаниями, и если бы он мог открыть дверь сам, то непременно это сделал бы (в таком случае $D_i^{Z=0} = 1$). Однако если хозяева открывают дверь ($Z_i = 1$), то кот обычно теряет интерес и уходит по своим делам в противоположную сторону: $D_i(Z_i = 1) = D_i^{Z=1} = 0$.

Что касается оставшихся двух страт, у них потенциальные воздействия являются одинаковыми вне зависимости от значений инструмента: $D^{Z=1} = D^{Z=0} = 1$ для *всегда за* (грубый, но показательный пример — те, кто начинает пить антибиотики при малейших симптомах ОРВИ и продолжает это делать, даже если врач прямо запрещает) и $D^{Z=1} = D^{Z=0} = 0$ для *всегда против* (те, кто принципиально отказывается ставить себе или своим детям прививку от некоторого заболевания, даже если врач рекомендует это сделать).

Таким образом, LATE–IV может быть неформально определен как средняя разница потенциальных исходов для тех наблюдений, которые следуют исходным предписаниям. Основная содержательная проблема с использованием LATE–IV на практике заключается в том, что мы обычно не можем определить, какие наблюдения являются *послушными*, а какие — нет, так как мы наблюдаем только одно (реализовавшееся) потенциальное воздействие. Иначе говоря, мы не можем понять, для кого именно целевая величина является репрезентативной. Впрочем, при определенных допущениях можно оценить долю таковых в выборке как $E(D_i|Z_i = 1) - E_i(D_i|Z_i = 0)$ и даже характеризовать основные моменты распределения наблюдаемых ковариат для этой страты [57; 61].

Следует также отметить, что инструментальная оценка может интерпретироваться в терминах LATE–IV только в том случае, если выполняется допущение о монотонности воздействия Z на D : $D_i^{Z=1} \geq D_i^{Z=0} \forall i$, т. е. при увеличении значения инструмента вероятность воздействия растёт (либо уменьшается) или остаётся неизменной для каждого наблюдения. В противном случае LATE–IV смешивает в неизвестной пропорции эффекты воздействия для разных страт. Это допущение является правдоподобным в ситуации *одностороннего непослушания* (англ. *ones-sided non-compliance*), когда участники исследования могут не соблюдать предписания, если они находятся в группе воздей-

ствия (например, не пить выданные врачом таблетки), но не в контрольной группе (у них нет доступа к этим таблеткам либо они вообще не знают о том, что пьют плацебо — как в двойном слепом эксперименте). Во многих полевых экспериментах социально-политической направленности, однако, существуют трудно поддающиеся контролю возможности для *двусторонне-го* (англ. *two-sided*) непослушания, в силу чего они редко способны дать надёжные оценки LATE–IV — не говоря уже об обсервационных исследованиях, в которых инструмент не рандомизируется в явном виде, а подбирается на основе здравого смысла и соответствия популярным, но при этом косвенным статистическим критериям валидности.

В случае с порядковыми и интервальными тритментами и/или многозначными и интервальными инструментами сохраняется базовая трактовка LATE–IV как среднего эффекта воздействия для наблюдений, у которых при изменении Z монотонно изменяется D . Впрочем, математическая и, как следствие, содержательная интерпретации существенно усложняются. По причине излишней громоздкости соответствующих формул я не привожу их здесь и отсылаю интересующихся читателей к более фундаментальным источникам [57, раздел 4.5.3; см. также 62; 63].

3.2 ИТТ

Ещё один целевой эффект, который можно оценить в инструментальных дизайнах и экспериментах с непослушанием (в том числе двусторонним) и который представляет практический интерес, — это так называемый эффект *намерения воздействовать* (англ. *intention-to-treat effect* — ИТТ): $\tau_{ITT} = E(Y_i^{Z=1} - Y_i^{Z=0})$. Следует прямо обозначить, что ИТТ не является локальным эффектом воздействия. Этот эстиманд показывает, как в среднем различаются исходы участников, включённых в номинальные группы воздействия и контроля, — независимо от того, было ли

оказано в реальности воздействие на конкретного участника (или не оказано, если речь о $Z_i = 0$). Если есть основания полагать, что доля *непослушных* является небольшой (и останется таковой при полномасштабной имплементации целевой политики), то такой эффект имеет прикладной смысл как более-менее информативная аппроксимация АТЕ. Однако в общем случае следует помнить, что ИТТ показывает эффект не самого воздействия (например, *фактически выпитого лекарства*), а его формального назначения (*выписанного рецепта на лекарство*), что не одно и то же.

С другой стороны, на практике назначенное воздействие (например, социальные выплаты или льготы) не всегда доходит до адресата. Так как ИТТ по определению учитывает возможность расхождения номинального и реального статуса воздействия, он позволяет получить представление о среднем эффекте политики в том — заведомо несовершенном — виде, в котором она будет осуществляться в реальности. В этом плане ИТТ может быть даже более полезным для лиц, принимающих решения, чем АТЕ, характеризующий эффект политики в идеальных условиях, которые вряд ли получится воспроизвести при масштабировании за рамки экспериментальной выборки.

3.3 LATE в разрывном дизайне

Разрывный дизайн обычно используется в тех случаях, когда имеется заранее заданный и определяемый объективными критериями порог, регулирующий назначение воздействия. Например, проходной балл на вступительных экзаменах в университет (набрал w баллов или больше — поступил; меньше — не поступил); нижняя граница средней оценки, при которой студент может претендовать на стипендию; географическая близость к границам исторического распространения какой-либо социальной практики (крепостное право или рабство);

разность полученных на выборах голосов между проигравшим и победившим кандидатом и т.д. [64].

Обозначим признак, на основании которого назначается воздействие, т. н. *классифицирующую* переменную (англ. *forcing/running variable*), как W , а пороговое значение — w_c . В разрывном дизайне целевым эффектом выступает условное математическое ожидание разницы потенциальных исходов под воздействием и без для $W = w_c$: $\tau_{LATE-RD} = E(Y_i^{D=1} - Y_i^{D=0} | W = w_c)$. Обычно этот эстиманд рассчитывается на основе экстраполяции $\hat{f}(Y_i^{D=0} | W_i = w)$ и $\hat{f}(Y_i^{D=1} | W_i = w)$, оцененных в регионах $[h^-, w_c)$ и $[w_c; h^+]$ соответственно. Предполагается, что в окрестностях порога наблюдения распределены между группами воздействия и контроля (квази)случайно, и поэтому разницу в исходах можно атрибутировать исключительно разнице статусов воздействия²⁶.

Вопрос о том, что такое «окрестности порога», является дискуссионным, и универсального критерия здесь нет. Иногда интервал $[h^-; h^+]$ может быть довольно большим в сравнении с областью определения W , однако целевой эффект в РД в любом случае является фундаментально локальным и характеризует результат воздействия только в пределах этого интервала. Это можно проиллюстрировать следующим образом. Предположим, что оценивается эффект поступления в университет по отношению к доходу индивида спустя 25 лет. Если взять круглого отличника, набравшего 100% возможных баллов на экзамене, и официально запретить ему получать высшее образование, то вряд ли кого-то удивит, что его будущий доход все равно

²⁶ Это один из вариантов теоретического обоснования разрывного дизайна, известный как парадигма локальной рандомизации (англ. *local randomization framework* [65]). Другой, более популярный (и, на мой взгляд, более тяжелый для объяснения), основывается на предположении о гладкости и непрерывности условных (по W) функций потенциальных исходов в окрестностях w_c .

окажется довольно высоким. Иначе говоря, для него контрфактуальная разница $E(Y_i^{D=1} - Y_i^{D=0} | W = w_{max})$ с высокой вероятностью будет меньше таковой у абитуриента, набравшего ровно w_c баллов. С другой стороны, если взять абстрактного круглого двоечника, набравшего ноль баллов на вступительном экзамене, и зачислить в обход тех, кто сдал вступительные лучше, то вряд ли он «доживёт» до конца первой сессии. Его шансы на успешную карьеру с высоким доходом изначально малы в обоих контрфактуальных сценариях, и ожидаемый эффект поступления в его случае также представляется небольшим.

С формальной точки зрения, разрывный дизайн позволяет (при условии, что выполняются необходимые допущения) оценить эффект поступления для абитуриентов, обладающих примерно одинаковым уровнем знаний, таланта, работоспособности, семейного социального капитала и других вещей, которые могут одновременно влиять на успехи в учёбе и на будущие доходы. В рамках выбранного примера РД полагается на допущение о том, что те, кто набрал $[w_c; h^+]$ баллов, ничем принципиально не отличаются от тех, кто набрал $[h^-, w_c)$ баллов, а неудача последних с поступлением является именно что *неудачей*, то есть вызвана идиосинкратическими случайными факторами (погода в день экзамена; случайная простуда или отравление; решил перевести бабушку через дорогу, началась перестрелка между полицией и бандитами, бабушка испугалась, упала всей массой студенту на палец и сломала его), а не систематическими различиями в успеваемости или умении зарабатывать деньги. Соответственно, в РД целевой эффект репрезентативен только для наблюдений, попадающих в интервал $[h^-; h^+]$. Более того, чем ближе значение W для наблюдения i к внешним границам этого интервала, тем в меньшей степени можно быть уверенным в том, что целевой эстиманд корректно характеризует, как попадание i в группу воздействия скажется на Y_i .

3.4 Выборочные и достижимые эффекты воздействия

Выше было обозначено, что для получения оценки АТТ (или АТС) часто используется синтетическая группа сравнения, которая формируется за счёт фильтрации группы контроля (для АТС — воздействия), т. е. отсеивания наблюдений, которые не похожи по своим измеренным характеристикам (ковариатам) X на наблюдения под воздействием и потому не могут рассматриваться как надежные контрфактуальные пары, позволяющие восстановить пропущенные потенциальные исходы. При этом некоторые наблюдения в группе воздействия тоже могут оказаться «выбросами», для которых невозможно найти подходящую пару. В таком случае ничего не остаётся, кроме как не учитывать такие «экстремальные» наблюдения в дальнейших расчётах. Однако у этого решения есть важное нежелательное последствие: получаемая оценка не является репрезентативной относительно АТТ, так как некоторые наблюдения под воздействием целенаправленно исключаются из анализа, причём не случайным образом, а согласно значениям X .

Впрочем, эту проблему можно частично обойти, переопределив эстиманд *post hoc* и задав в качестве цели вместо АТТ так называемый *достижимый* (или *реалистичный*) *средний эффект воздействия для группы воздействия в выборке* — FSATT (англ. *feasible sample average treatment effect on the treated*): $\tau_{FSATT} = E(Y_i^{D=1} - Y_i^{D=0} | D_i = 1, M_i = 1)$, где M_i — это индикаторная переменная, равная 1, если i -тое наблюдение было отобрано по результатам мэтчинга, и 0 — если нет [55]. Данный эстиманд имеет ограниченную информативность, так как он, строго говоря, характеризует эффект воздействия (а) исключительно в доступной исследователю выборке и только для (б) наблюдений из группы воздействия, которым (в) удалось подобрать подходящую пару в группе контроля. Возможность его ге-

нерализации на другие наблюдения (особенно со значениями X_i , отличающимися в среднем от таковых в выборке) из целевой популяции в общем остается сомнительной.

С другой стороны, и в обсервационных исследованиях (не только тех, в которых используется мэтчинг), и в экспериментах даже исходные выборки зачастую отклоняются от генеральной совокупности по многим параметрам, поэтому проблема репрезентативности полученных оценок на практике может стать актуальной для любого казуального дизайна. Наиболее очевидным решением, основанном на принципе «лучше *вытащить из данных хотя бы что-то*, чем вообще ничего», является назначение целевым эстимандом как среднего эффекта воздействия в выборке (англ. *sample average treatment effect* — SATE): $\tau_{SATE} = E(Y_i^{D=1} - Y_i^{D=0} | S_i = 1)$, где S_i — это индикаторная переменная, отражающая факт попадания в выборку (SATТ, SATC и другие выборочные версии рассмотренных выше эстимандов задаются аналогично). Однако при добавлении буквы S в аббревиатуру, обозначающую эстиманд, необходимо четко указывать на интерпретационные ограничения, из этого происходящие (см. абзац про FSATТ).

Популярным частным случаем реалистичного эстиманда является средний эффект воздействия в *пересечении* (англ. *overlap*) или в *клиническом равновесии* (англ. *equipoise* [66]) — АТО. Он характеризует результат воздействия в той ограниченной совокупности наблюдений/индивидов, для которой конкретное воздействие имеет теоретический смысл, но его эмпирическая польза неочевидна [8].

Вряд ли нужно предлагать безусловный базовый доход Илону Маску или Роману Абрамовичу. Однако для людей, живущих у или за гранью порога бедности, назначение ББД может иметь положительный эффект (например, повысить их шансы найти работу или улучшить ментальное состояние). А может и не иметь — собственно, потенциальная величина такого

эффекта как раз и представляет основной интерес для лиц, принимающих решение о запуске соответствующей программы. Точно так же не стоит делать пересадку сердца человеку, который впервые в жизни пришел к врачу с жалобами на давление 131 на 69, или удалять всю челюсть при малейших признаках кариеса на одном из зубов.

Эти примеры могут показаться гротескными, но они выбраны намеренно, чтобы проиллюстрировать основную идею, стоящую за использованием АТО: большая часть воздействий или политик предназначены не для всей ГС, а для определенных её подмножеств, характеризующихся теми или иными признаками, мотивирующими необходимость вмешательства (бедность, серьёзные проблемы с сердцем или зубами и т. д.). Поэтому оценивать эффективность этих воздействий и политик надо в первую очередь для целевых подмножеств.

Если воспользоваться менее экстремальным примером, то препараты от рака, предназначенные для применения на ранних этапах заболевания, не следует испытывать на совершенно здоровых людях или на пациентах с IV стадией. Обычно их проверяют на тех, кому уже диагностирована болезнь, но при этом они имеют шанс вылечиться. Такая ситуация хорошо показывает основную особенность АТО — его зачастую используют для оценки сравнительной эффективности конкурирующих терапий или политик (новой и старой или альтернативных новых). Целевой совокупностью в данном случае выступают наблюдения, которые имеют примерно равные — то есть близкие к 0,5 — предсказанные вероятности попадания как в группу воздействия, так и в группу контроля. В обсервационных исследованиях эту популяцию (точнее, выборку из неё) часто определяют путем наложения распределений мер склонности в обеих группах друг на друга

и последующего выявления области их *пересечения*²⁷ (отсюда *O* — *overlap* — в аббревиатуре).

3.5 CATE

Почти все рассмотренные выше эстиманды (и бо́льшая часть обсуждаемых далее) являются частными случаями условных средних эффектов воздействия. В LATE–IV условием является принадлежность к группе «послушных»; в разрывном дизайне — близость к порогу; в выборочных эстимандах — факт попадания в выборку (в том числе по результатам процедур мэтчинга или взвешивания). Однако рассмотренными примерами семейство CATE не ограничивается. В общем виде данный тип эстимандов можно определить как $\tau_x = E(Y_i^{D=1} - Y_i^{D=0} | X_i = x)$, где x — это либо значение одной конкретной ковариаты, либо вектор индивидуальных значений всех доступных ковариат для i -го наблюдения.

В первом случае обычно имеется в виду принадлежность наблюдения i к какой-то группе или страте. Соответственно, CATE²⁸ для заданного x характеризует средний ожидаемый размер эффекта для целевой страты. Как видно из формального

²⁷ Так как в большинстве обсервационных сценариев (в частности, в медицине) назначение воздействия происходит неслучайным образом и сильно зависит от известных индивидуальных характеристик, распределения мер склонности в группе воздействия зачастую сильно скошено влево (т. е. пик смещён к единице), а в контрольной группе — вправо (пик находится около нуля) [8, р. 16, рисунок В]. Поэтому область их пересечения включает существенно меньше наблюдений, чем в целом доступно в данных, в силу чего АТО не является тождественным АТЕ. При этом она не обязательно должна ограничиваться исключительно наблюдениями с $e = 0.5$: допустимый размах её границ определяется контекстом и целями исследования.

²⁸ τ_x может быть не только CATE, но и CATT или каким-то другим эстимандом, в том числе локальным.

определения, АТЕ и другие типы каузальных эстимандов не предполагают константного характера воздействия. Напротив, все они усредняют индивидуальные эффекты, которые сами по себе могут варьироваться в произвольной степени вплоть до того, что знак эффекта различается между конкретными наблюдениями. Впрочем, часто практический интерес представляет гетерогенность несколько меньшего масштаба, а именно зависимость целевого эффекта от значений некоторой ковариаты X (или комбинации нескольких X)²⁹. Как различается эффективность одного и того же лекарства для молодых и пожилых пациентов? Одинаковым ли образом введение материального капитала повлияло на рождаемость в богатых и бедных регионах? В регионах с высокой и с низкой рождаемостью в предшествующий период? Использование САТЕ в качестве целевого эффекта позволяет операционализировать подобные запросы к данным [67].

Во втором случае речь идет, с одной стороны, об ожидаемом эффекте воздействия для всех наблюдений, которые характеризуются определенным профилем значений ковариат (женщина 1997 года рождения с высшим образованием, незамужняя, таких-то политических взглядов; далее идёт полная расшифровка генома; продолжать можно до бесконечности), т. е., строго говоря, также о стратоспецифичном эффекте. С другой стороны, в силу значительного количества учитываемых признаков и приистекающего из этого малого, в том числе единичного и даже нулевого размера некоторых страт, такой вариант САТЕ

²⁹ В регрессионном анализе такая переменная часто называется *модератором*, а само явление вариативности размера эффекта одной переменной в зависимости от значений другой — *модерацией* или *взаимодействием* (в отечественной литературе также встречается термин *эффект взаимодействия*, а в англоязычной — *interaction effect*).

может рассматриваться как попытка аппроксимации индивидуального эффекта воздействия τ_i для отдельно взятых i .

Последний эстиманд имеет несомненную прикладную ценность: любой врач, выписывая лекарство или предлагая терапию, хотел бы знать, как эти лекарство/терапия подействуют на пациента, который сидит перед ним здесь и сейчас, со всеми его уникальными особенностями, а не на условного «среднего пациента по больнице». Это справедливо и для любой другой сферы, которую можно анализировать с помощью методов статистического причинного-следственного вывода — от имплементации социальных программ до индивидуализации пользовательских интерфейсов в онлайн-индустрии. Однако оценка таких детализированных эстимандов предъявляет существенно более высокие требования к данным и в общем случае является нетривиальной (зачастую попросту неразрешимой) задачей. Тем не менее прогресс в технологиях сбора т. н. «больших» данных и их предиктивного анализа с помощью методов машинного обучения в последние годы поспособствовал активному развитию литературы по гетерогенным эффектам воздействия, в которой основными целевыми эстимандами как раз выступают CATE второго типа (например, [68]; краткий, но полезный и насыщенный обзор даётся в [69, раздел 3]).

3.6 Каузальные взаимодействия

Отдельный интерес представляют так называемые каузальные взаимодействия (англ. *causal interactions*). Такое взаимодействие имеет место, когда $E(Y_i^{D_1=1, D_2=1}) - E(Y_i^{D_1=0, D_2=1}) \neq E(Y_i^{D_1=1, D_2=0}) - E(Y_i^{D_1=0, D_2=0})$. Содержательно оно может возникать в многофакторных экспериментах, например в комбинационных (англ. *conjoint*), являющихся популярными инструментами изучения потребительских и политических предпочтений [70]. В обсервационных исследованиях каузальные

взаимодействия тоже вероятны: так, в эпидемиологии типичным является сценарий, когда эффективность какой-либо терапии или риск развития заболевания из-за негативного воздействия внешнего фактора (курения в случае рака легких) могут увеличиваться (снижаться) для людей с наличием определенного генетического варианта, который сам по себе причинно-следственным образом связан с целевым исходом [71].

Для квантификации каузальных воздействий могут использоваться несколько теоретических эстимандов. Во-первых, *средний эффект комбинации* (англ. *average combination effect* — ACE; однако я его обозначаю как *Comb*, т. к. аббревиатура ACE часто используется для сокращения синонимичного АТЕ термина *average causal effect*): $\tau_{Comb} = E(Y_i^{D_1=1, D_2=1}) - E(Y_i^{D_1=0, D_2=0})$. Он показывает ожидаемую разницу потенциальных исходов в случае одновременного воздействия двух факторов против отсутствия какого-либо воздействия вообще. Содержательный смысл этого эффекта можно проиллюстрировать в контексте популярного метода изучения дискриминации на рынке труда посредством рассылки фиктивных резюме [72; 73]. Если в резюме случайным образом варьируют две переменные — пол (мужчина/женщина) и этничность (титульная нация/меньшинство), то средний эффект комбинации отражает контрфактуальную разницу в вероятности отклика для резюме, поданных мужчинами, принадлежащими к доминирующему этносу, и женщинами, принадлежащими к этническому меньшинству.

Во-вторых, *средний предельный эффект воздействия* одного из факторов (англ. *average marginal effect* — AME), усреднённый по значениям другого фактора: $\tau_{AME} = \int E(Y_i^{D_1=1, D_2=d} - Y_i^{D_1=0, D_2=d}) dF(D_2)$. В рамках того же примера это средневзвешенный по полу эффект принадлежности к титульной нации.

В-третьих, *средний эффект взаимодействия* (англ. *average interaction effect* — AIE): $\tau_{AIE} = E(Y_i^{D_1=1, D_2=1} - Y_i^{D_1=1, D_2=0} - Y_i^{D_1=0, D_2=1} + Y_i^{D_1=0, D_2=0})$. Этот эстиманд может интерпретироваться как разность в АТЕ пола между представителями титульного этноса и меньшинства (или АТЕ этничности между мужчинами и женщинами).

Наконец, *средний предельный эффект* взаимодействия (англ. *average marginal interaction effect* — AMIE): $\tau_{AMIE} = \tau_{Comb} - \tau_{AME, D_1} - \tau_{AME, D_2}$. Данная величина показывает, насколько ожидаемый эффект от комбинации двух факторов (женщина & этническое меньшинство vs мужчина & большинство) отличается от суммы частных эффектов (женщина vs мужчина + меньшинство vs большинство). Содержательно она близка AIE: обе величины являются линейными функциями друг друга. Однако AMIE инвариантен к выбору базовой комбинации ($D_1 = 0, D_2 = 0$) и обладает рядом полезных свойств при работе со взаимодействиями высоких порядков [74].

3.7 Опосредованные эффекты

Социальные и социально-биологические процессы являются комплексными, и, как правило, изменение D не транслируется напрямую в изменение Y , а влияет опосредованно, через воздействие на так называемые переменные-медиаторы (M). Цепь промежуточных эффектов $D \rightarrow M_1 \rightarrow \dots \rightarrow M_k \rightarrow Y$ представляет собой каузальный механизм, связывающий целевые D и Y , причём зачастую таких механизмов, или путей, может быть несколько. Понимание механизмов, стоящих за конкретным воздействием, способствует успешной имплементации соответствующей политики и поэтому представляет отдельный исследовательский интерес. Основным статистическим инструментом для изучения механизмов, в том числе каузальных, является

анализ опосредованных эффектов, или медиационный анализ (англ. *mediation analysis*) [75; 76].

Как и в случае инструментальных переменных, для представления опосредованных эффектов в терминах потенциальных исходов требуется ввести дополнительные величины, а именно потенциальные медиаторы. Для упрощения нотации в этом разделе используются обозначения $Y_i(d) = Y_i^{D=d}$ и $M_i(d) = M_i^{D=d}$. Общий (англ. *total*) индивидуальный эффект воздействия — это величина $\tau_i = Y_i\{1, M_i(1)\} - Y_i\{0, M_i(0)\}$. Он раскладывается на *опосредованное воздействие* $\delta_i(d) = Y_i\{d, M_i(1) - Y_i(d, M_i(0))\}$ и *прямое воздействие* $\xi_i(d) = Y_i\{1, M_i(d) - Y_i(0, M_i(d))\}$ (в обоих выражениях $d = 0, 1$): $\tau_i = \delta_i(d) + \xi_i(1 - d)$.

$\delta_i(d)$ показывает, насколько изменится Y_i , если $M_i(1)$ заменить на $M_i(0)$ для заданного значения D , которое остаётся *фиксированным*, т. е. не меняется. Другими словами, $\delta_i(d)$ характеризует отклик Y_i исключительно на изменение в M_i из-за воздействия D_i . В свою очередь, $\xi_i(d)$ показывает, насколько изменится Y_i , если изменить D_i для фиксированного значения потенциального медиатора $M_i(d)$. Проще говоря, эта величина описывает, насколько сильно влияет D на Y , если исключить механизм $D \rightarrow M \rightarrow Y$. Соответствующие усредненные эстиманды называются *средний эффект опосредованного воздействия* (англ. *average causal mediation effect* — ACME) и *средний эффект прямого воздействия* (англ. *average direct effect* — ADE). Необходимо обратить внимание на то, что и ACME, и ADE можно задать как для группы воздействия и для группы контроля по отдельности, так и усреднить.

Рассмотрим следующий пример. Известно, что бедность во многом наследуется: если человек рос в семье с низким доходом, он с большой вероятностью будет бедным и в последующих фазах жизненного цикла. Одним из медиаторов здесь выступает образование [77]. Для простоты предположим, что опыт бедности в детстве является дихотомической переменной:

$D_i = 1$, если имеется, и $D_i = 0$, если нет. В таком случае АСМЕ для $d = 1$ показывает, насколько в среднем теряют в доходах во взрослом возрасте люди из бедной семьи по сравнению со сценарием, в котором их потенциальный уровень образования в условиях бедности $M_i(1)$ заменяется потенциальным уровнем образования $M_i(0)$, который они получили бы, если росли в богатой семье. Соответственно, АДЕ для $d = 1$ показывает, насколько в среднем теряют в доходах во взрослом возрасте люди из бедной семьи по сравнению с контрфактуальным сценарием, в котором они бы росли в богатой семье, но с тем же уровнем образования, $M_i(1)$, что в бедной. АСМЕ и АДЕ для $d = 0$ определяются аналогично.

4. Расширения

4.1 Y не является интервальной переменной

Во многих приложениях целевая зависимая переменная не является нормально распределённой непрерывной величиной. Например, часто интерес представляют эффекты воздействия по отношению к дихотомическим признакам: способствует ли прохождение программы переобучения успешному поиску работы (да/нет); повышает ли применение нового метода лечения рака шансы пациента выжить и т. д. Однако и для дихотомических, и для любых других возможных шкал сохраняется общая логика задания целевых эстимандов в терминах контрастов потенциальных исходов под воздействием и без.

Для предложенных примеров стандартный АТЕ является вполне осмысленной величиной, которую можно интерпретировать как разницу между двумя контрфактуальными сценариями в долях тех, кто нашел работу или выжил. В эпидемиологии этот показатель известен как каузальная разница рисков (англ. *causal risk difference*): $\tau_{RD} = P(Y^{D=1} = 1) - P(Y^{D=0} = 1)$.

Впрочем, для дихотомических Y целевой эстиманд может быть представлен и с помощью других метрик, таких как каузальное отношение³⁰ рисков (англ. *risk ratio*), $\tau_{RR} = \frac{P(Y^{D=1}=1)}{P(Y^{D=0}=1)}$ (во сколько раз вероятность найти работу/выжить под воздействием превышает аналогичный показатель без воздействия) или каузальное отношение шансов (англ. *odds ratio*), $\tau_{OR} = \frac{P(Y^{D=1}=1)}{P(Y^{D=1}=0)} / \frac{P(Y^{D=0}=1)}{P(Y^{D=0}=0)}$ (то же самое, только для шансов вместо вероятностей).

Нужно отметить, что RD, RR и OR — это разные *метрики*, или *шкалы*, величины целевого эффекта (ATE или любой альтернативы), а не самостоятельные эстиманды. Поэтому, в частности, в соответствующих формулах отсутствуют нижние индексы: вероятности $P(Y^{D=d} = y)$ являются усредненными, а не индивидуальными контрфактуальными величинами.

Ещё один важный нюанс при работе с дихотомическими Y — это различие между разложимыми и неразложимыми (англ. *collapsible* и *non-collapsible*) метриками эффекта. Первые могут быть представлены как взвешенная сумма стратоспецифичных (т. е. условных) эффектов воздействия (среди женщин и среди мужчин; среди разных возрастных групп; для разных стадий болезни) в условиях гетерогенности эффекта, тогда как вторые — нет, в силу чего порой возникают контринтуитивные результаты, согласно которым основной эффект имеет меньшую величину, чем любая из его стратоспецифичных составляющих.

Кроме того, для неразложимых метрик безусловные (англ. *marginal*) и условные (англ. *conditional*) эстиманды не могут

³⁰ Хотя в настоящей работе мультипликативная шкала эффекта (во сколько раз отличаются ожидаемые потенциальные исходы между условиями воздействия) обсуждается в контексте дихотомических зависимых переменных, её можно использовать и с интервальными Y . Впрочем, такой подход встречается редко, и аддитивная шкала (на сколько) является выбором по умолчанию.

быть оценены с использованием одного и того же подхода, что имеет прямые практические последствия с точки зрения выбора целевого эффекта воздействия. Так, если для оценки τ_{OR} используется поправка на наблюдаемые ковариаты X посредством стандартной практики добавления контрольных переменных в логистическую регрессию $Y \sim D$, то даже при выполнении допущений о взаимозаменяемости и пересечении коэффициент при D характеризует не τ_{OR} , а условный $\tau_{OR}(X)$, в качестве составных элементов которого используются условные контрфактуальные вероятности $P(Y^{D=d} = y | X)$. $\tau_{OR}(X)$ — это средний эффект воздействия для конкретных комбинаций переменных в множестве X ; он является разным для разных x [78; 79]. Кроме того, для множества X' популяционный эффект по шкале отношения шансов будет отличаться от такового для X .

Разложимость зависит от многих факторов: одна и та же метрика может быть разложимой (хотя бы приблизительно) в одном контексте и неразложимой в другом. В большинстве случаев RD и RR — это разложимые метрики, а OR или HR (англ. *hazard ratio* — основная метрика эффекта в анализе выживаемости) — неразложимые [32, р. 56; 80; 81]. В этой связи задание целевого эффекта по отношению к дихотомическим Y с помощью аддитивной шкалы представляется менее подверженной интерпретационным ошибкам практикой. Подробное и одновременно доступное обсуждение феномена (не)разложимости и его концептуальных и практических импликаций см. в [78].

В продолжение темы метрик эффекта стоит заметить, что эстиманд можно формулировать не только в терминах контрастов математических ожиданий, но и медиан или других квантилей, и даже кумулятивных функций распределения гетерогенного эффекта воздействия в целом [82; 83]. Информацию о возможных целевых эффектах при работе с порядковыми Y можно найти, например, в [84; 85].

4.2 D имеет больше двух категорий

Если D является дискретной переменной, которая имеет больше двух значений, как в экспериментах, где сравниваются два разных воздействия (например, два перспективных варианта лечения одного и того же заболевания) и контрольное условие (принятый метод лечения), то задаются попарные контрфактуальные контрасты (ATE или любой другой, представляющий интерес), характеризующие эффекты каждого из воздействий относительно статус-кво. Можно также напрямую сравнить средние потенциальные исходы под обоими воздействиями: $E(Y_i^{d1}) - E(Y_i^{d2})$. Эта величина будет отражать ожидаемую разницу в эффективности $d1$ и $d2$.

Чем больше значений k принимает D , тем больше возможных контрфактуальных контрастов: $k(k - 1)/2$. Как следствие, и интерпретация множества целевых эффектов, и их эмпирическая оценка на конечной выборке становятся проблематичными. Поэтому, исходя из соображений практической целесообразности, в частности баланса временных, финансовых и трудовых затрат с одной стороны и надёжности полученных результатов — с другой, лучше избегать таких сценариев и ограничиваться небольшим числом сравниваемых альтернатив.

Если D представляет собой набор упорядоченных значений, то можно задать частные контрасты конкретных уровней $D = d$ против базового уровня, $\tau_d = E(Y_i^{D=d} - Y_i^{D=0})$, показывающие ожидаемый выигрыш (ухудшение) по сравнению с полным отсутствием воздействия (или его минимальным объёмом): насколько различается вероятность развития рака лёгких, если выкуривать d сигарет в день или не курить вообще. Другой набор релевантных для принятия решений эстимандов включает контрасты более высоких уровней против более низких, $\tau_{d'-d} = E(Y_i^{D=d'} - Y_i^{D=d})$, $d' > d$: эффект 15 сигарет в день против 10 сигарет. Частным вариантом являются контра-

сты соседних уровней, $\tau_{d+1} = E(Y_i^{D=d+1} - Y_i^{D=d})$: эффект двух пачек сигарет против одной пачки или эффект получения среднего общего образования против основного общего образования по отношению к доходам/количеству детей в возрасте 40 лет. Эстиманды, отражающие эффекты прироста на одну единицу для разных уровней d можно затем усреднить по уровням d . Однако репрезентативность этой величины относительно её частных составляющих вызывает вопросы в тех случаях, когда размер эффекта не является постоянным по d [62; 63; 86].

Наконец, если D является полноценной интервальной переменной или может рассматриваться в качестве такой (например, зарплата), то основным эстимандом является так называемая *доза — отклик* (англ. *dose — response*; далее — ФДО), $E(Y_i^{D=d})$ [87; 88]. Эта функция характеризует ожидаемые (усреднённые) значения потенциальных исходов для различных значений воздействующей переменной: каким будет уровень удовлетворенности жизнью при зарплате r рублей? $2r$ рублей? $5r$ рублей? Зная ФДО, можно ответить на вопросы типа «насколько вырастет в среднем удовлетворенность жизнью при повышении зарплаты с нуля или с r рублей до $2r$ рублей» с помощью эстимандов τ_d и $\tau_{d'-d}$, аналогичных таковым для порядковых D . В качестве дополнительного целевого параметра для подобных контекстов в литературе также выделяют производную ФДО в точке d : $\tau_{ACR_d} = \frac{\partial E(Y_i^{D=d})}{\partial d}$ (англ. *average causal response* — ACR [86]), характеризующую предельный эффект прироста воздействующей переменной для выбранного уровня.

Так как число возможных d для непрерывных D является бесконечным, интуитивно привлекательной представляется идея обобщить частные эффекты с помощью некоторой суммирующей статистики, такой как $\bar{\tau}_d = E(\tau_d | D > 0)$

или $\bar{\tau}_{ACR_d} = E(\tau_{ACR_d} | D > 0)$. Подобные эстиманды, однако, будут информативны только в том случае, если ФДО является линейной или хотя бы приблизительно линейной. На практике это не всегда так, поэтому для интервальных D всё же в качестве эстиманда лучше фокусироваться именно на ФДО в целом (см. пример в [86]). При необходимости затем можно выбрать референтные значения d и d' , исходя из контекста конкретной задачи.

Также необходимо заметить, что в общем случае для многозначных, порядковых и интервальных D вполне осмысленными являются все те модификации АТЕ, которые ранее рассматривались для дихотомических воздействующих переменных: эффекты для группы воздействия, локальные и условные эффекты и т. д. В частности, доступное обсуждение интерпретации LATE-IV для порядковых и интервальных воздействий можно найти в разделе 4.5.3 «Почти безвредной эконометрики» [57].

4.3 Целевые эффекты в лонгитюдных дизайнах

Данные, имеющие временное измерение, предоставляют широкие возможности как для описательного, так и для доказательного количественного анализа. Обратной стороной присутствующего им изобилия информации является не меньшее изобилие возможных целевых эффектов.

Рассмотрим наиболее популярный каузальный дизайн для лонгитюдных исследований — метод *разность разностей* (англ. *difference-in-differences* [89]), или РР. В самой простом РР-сценарии с двумя временными срезами, t_0 (до воздействия) и t_1 (после воздействия), целевым эффектом выступает АТТ³¹

³¹ АТТ является основным эстимандом и в других лонгитюдных каузальных дизайнах, таких как метод синтетических контролей или прерванные временные ряды.

для t_1 : $\tau_{DiD} = E(Y_{i,t1}^{D=1} - Y_{i,t1}^{D=0} | D_{i,t1} = 1)$ [57; 90; 91; 92; 93; 94]. Контрольные наблюдения в данном случае используются для восстановления информации о контрфактуальном тренде без воздействия для группы воздействия: предполагается, что без воздействия тренды в обеих группах были бы одинаковыми — даже если наблюдаемые средние значения Y в момент t_0 отличаются (т. н. допущение о *параллельных трендах* — англ. *parallel trends*). Эффект воздействия для контрольной группы, таким образом, выносится за скобки.

Если мы наблюдаем более одного среза после воздействия, то АТТ можно определить для каждого из них ($t_1, t_2, \dots$). Зачастую это имеет практический смысл: сравнивая оценки АТТ для разных периодов после воздействия, можно отследить, имеет ли оно накопительный эффект и насколько длительным последний является. Когда для принятия решения требуется единственная количественная величина, можно взять среднее значение по частным АТТ для отдельных срезов (при необходимости — с весами) или для конкретного среза, который представляется наиболее уместным с прикладной точки зрения.

Более сложная ситуация возникает, если воздействие для разных наблюдений происходит в разные моменты времени (т. н. называемое *рассеянное* или *разнесённое* [во времени — прим. авт.] получение воздействия — англ. *staggered treatment adoption*). Например, одни пользователи социальной сети могут переходить на премиум-подписку сразу после её введения, а другие — в какой-то произвольный момент в будущем. Из двух пациентов, наблюдающихся у одного терапевта, один сразу перейдёт на новую методику лечения, а другой — спустя несколько месяцев. Из двух семей, у которых дети родились в один день, одна использует сертификат на маткапитал сразу после его выдачи, а другая — через несколько лет. В таких сценариях можно разделить выборку на когорты g , объединяющие

наблюдения, для которых воздействие произошло в один и тот же момент t^g .

Соответственно, АТТ можно оценить для каждой уникальной пары $\{g, t\}$: $\tau_{DiD}(g, t) = E(Y_{i,t \geq 0}^{D=1} - Y_{i,t \geq 0}^{D=0} | G_i = g)$. Если g и t у нас много, таких частных величин тоже будет много. Это представляется полезным с точки зрения понимания масштабов гетерогенности целевого эффекта, но затрудняет интерпретацию и представление результатов ответственным лицам или внешней публике. Наиболее очевидным решением проблемы опять-таки является усреднение, но здесь возможны уже три варианта, а не один: усреднять по g , по t или по $\{g, t\}$ одновременно. В первом случае мы имеем дело со средним АТТ для выбранной когорты (англ. *cohort* АТТ) за все периоды после воздействия. Во втором более осмысленно учитывать не абсолютное время t , а период относительно начала воздействия l ($l = -1$ в период непосредственно перед воздействием; ..., $-3, -2$ в предыдущие периоды; 0 в период непосредственно после воздействия; $1, 2, 3$... в последующие периоды)³².

Если используется абсолютное время и воздействие начинается в разные t , то в один и тот же период на наблюдение i воздействие оказывается уже продолжительный срок, тогда как на наблюдение i' — совсем небольшой. Другими словами, усреднение происходит по когортам, в одной из которых пациенты принимают лекарство полгода, а в других — всего месяц или даже один день. Если используется относительное время, то соответствующие эстиманды показывают период-специфичные АТТ (которые часто обозначают как динамические эффекты воздействия — англ. *dynamic treatment effects* [91; 92]) для всех

³² Во многих источниках период непосредственно до воздействия индексируется как 0 , а период непосредственно после воздействия — как 1 . Интерпретация динамических АТТ от этого не меняется.

наблюдений, на которые воздействие оказывается в течение как минимум l периодов.

Третий тип агрегирования, самый детализированный — уже описанные выше эффекты для пар $\{g, t\}$ (англ. *group-time effects*), — позволяет отразить различия между когортами в динамике выраженности воздействия, а также может использоваться для диагностики ключевого для большинства методов каузального вывода в лонгитюдных сценариях предположения о параллельных трендах: если некоторые $ATT_{\{g,t\}}$ для $l_g < 0$ (т. е. периодов до воздействия для когорты g) значимо отличаются от нуля, это свидетельствует против указанного допущения.

Ситуация, когда воздействие может назначаться и затем отменяться, является ещё более сложной и поэтому здесь не рассматривается (см., например, [94]), равно как и интерпретация эффектов интервальных непрерывных воздействий в лонгитюдных контекстах [86].

Заключение

Цель данной работы состояла в том, чтобы предложить отечественному читателю обзор основных каузальных эстимандов, то есть контрфактуальных контрастов, позволяющих операционализировать в количественных терминах содержательные вопросы о том, какой эффект будет иметь некоторое воздействие, или изменение статус-кво, в конкретной прикладной сфере — будь то государственное управление, бизнес-процессы, медицинская практика или научный эксперимент.

Зная содержательный смысл и область генерализации различных эстимандов, аналитики, непосредственно работающие с данными, могут легко переводить результаты своей работы на язык, понятный принимающим решения лицам, равно как и наоборот — формализовывать запросы, поступающие от по-

следних, в терминах контрфактуальной модели Рубина (или Пёрла — в зависимости от личных пристрастий).

Разумеется, постановка содержательного вопроса и его формализация — это лишь первый шаг большого пути, который нужно пройти для получения надёжных статистических оценок эффективности той или иной политики. К сожалению, экономические, этические и другие ограничения зачастую приводят к тому, что не на каждый поставленный вопрос удаётся ответить.

Однако если вопрос даже не поставлен, то ответить на него невозможно в принципе — в том числе и понять, что имеющих-ся данных попросту недостаточно или они не соответствуют задаче. Дорога в тысячу ли начинается с одного шага, а доказательный статистический анализ (любой, не только каузальный) начинается с выбора эстиманда, который определяет последующие действия исследователя. Смею надеяться, что в этом отношении представленный обзор окажется полезным как тем читателям, кто в силу служебного положения задаёт каузальные вопросы, касающиеся последствий изменений действующей «политики», так и тем, кто ищет ответы на эти вопросы с помощью количественных инструментов.

ЛИТЕРАТУРА

1. Ениколопов Р. С.* Доказательная экономика развития: Нобелевская премия по экономике 2019 года // Вопросы экономики. 2020, № 1. С. 5–17. DOI: 10.32609/0042-8736-2020-1-5-17. EDN: WXVLIN.
2. Серебренников Д. Е., Кузьмина Ю. В. Полевые эксперименты и модель причинно-следственного вывода Дональда Рубина: обзор актуальных исследований // Экономическая социология. 2021, т. 22, № 4. С. 117–139. DOI: 10.17323/1726-3247-2021-4-117-139. EDN: HQERRJ.
3. Heckman J. J. Econometric causality // International statistical review. 2008, vol. 76, № 1. P. 1–27. DOI: 10.1111/j.1751-5823.2007.00024.x.

* Внесен в список иностранных агентов Минюста РФ.

4. Сонин К. И.* Вмesto лаборатории: анализ данных естественных экспериментов (Нобелевская премия по экономике 2021 года) // Вопросы экономики. 2022, № 1. С. 5–22. DOI: 10.32609/0042-8736-2022-1-5-22. EDN: MHOZVB.
5. *Hitchcock C.* Probabilistic causation // The Stanford encyclopedia of philosophy (Spring 2021 Edition) / Ed. By E. N. Zalta. 2021. URL: <https://plato.stanford.edu/archives/spr2021/entries/causation-probabilistic/> (дата обращения: 05.10.2025).
6. *Imbens G. W.* Causal inference in the social sciences // Annual review of statistics and its application. 2024, vol. 11. P. 123–152. DOI: 10.1146/annurev-statistics-033121-114601.
7. *Hassell H. J., Holbein J. B.* Navigating potential pitfalls in difference-in-differences designs: Reconciling conflicting findings on mass shootings' effect on electoral outcomes // American political science review. 2024, vol. 119, № 1. P. 240–260. DOI: 10.1017/S0003055424000108.
8. *Greifer N., Stuart E. A.* Choosing the estimand when matching or weighting in observational studies. arXiv:2106.10577 [Stat], June 2021. DOI: 10.48550/arXiv.2106.10577.
9. *Heiss A.* Demystifying causal inference estimands: ATE, ATT, and ATU // Andrew Heiss: [сайт]. 21.03.2024. DOI: 10.59350/c9z3a-rcq16. URL: <https://www.andrewheiss.com/blog/2024/03/21/demystifying-ate-att-atu/> (дата обращения: 05.10.2025).
10. *Barrett M., D'Agostino McGowan L., Gerke T.* Causal inference in R // R-Causal: [сайт]. 21.08.2025. URL: <https://www.r-causal.org/> (дата обращения: 05.10.2025).
11. *Lundberg I., Johnson R., Stewart B. M.* What is your estimand? Defining the target quantity connects statistical evidence to theory // American sociological review. 2021, vol. 86, № 3. P. 532–565. DOI: 10.1177/00031224211004187.
12. *Neyman (Splawa-Neyman) J.* On the application of probability theory to agricultural experiments. Essay on principles. Section 9 // Roczniki Nauk Rolniczych Tom X. 1923 [in Polish]; reprinted in Statistical science. 1990, vol. 5, № 4. P. 465–472 / Transl. by D. M. Dabrowska, T. P. Speed. URL: <https://www.jstor.org/stable/2245382> (дата обращения: 05.10.2025).
13. *Holland P. W.* Statistics and causal inference // Journal of the American statistical association. 1986, vol. 81, № 396. P. 945–960. DOI: 10.1080/01621459.1986.10478354.
14. *Rubin D. B.* [On the application of probability theory to agricultural experiments. Essay on principles. Section 9.] Comment: Neyman (1923) and causal

* Внесен в список иностранных агентов Минюста РФ.

inference in experiments and observational studies // *Statistical science*. 1990, vol. 5, № 4. P. 472–480. URL: <https://www.jstor.org/stable/2245383> (дата обращения: 05.10.2025).

15. *Angrist J. D., Imbens G. W., Rubin D. B.* Identification of causal effects using instrumental variables: Rejoinder // *Journal of the American statistical association*. 1996, vol. 91, № 434. P. 468–472. DOI: 10.2307/2291634.

16. *Rubin D. B.* Causal inference using potential outcomes: Design, modeling, decisions // *Journal of the American statistical association*. 2005, vol. 100, № 469. P. 322–331. DOI: 10.1198/016214504000001880.

17. *Heckman J. J.* Identification of causal effects using instrumental variables: Comment // *Journal of the American statistical association*. 1996, vol. 91, № 434. P. 459–462. DOI: 10.2307/2291631.

18. *Pearl J.* Causal inference: History, perspectives, adventures, and unification (An interview with Judea Pearl) // *Observational studies*. 2022, vol. 8, № 2. P. 23–36. DOI: 10.1353/obs.2022.0007.

19. *Rubin D. B.* Matching to remove bias in observational studies // *Biometrics*. 1973, vol. 29, № 1. P. 159–183. DOI: 10.2307/2529684.

20. *Rubin D. B.* The use of matched sampling and regression adjustment to remove bias in observational studies // *Biometrics*. 1973, vol. 29, № 1. P. 185–203. DOI: 10.2307/2529685.

21. *Rubin D. B.* Estimating causal effects of treatments in randomized and nonrandomized studies // *Journal of educational psychology*. 1974, vol. 66, № 5. P. 688–701. DOI: 10.1037/h0037350.

22. *Rubin D. B.* Assignment to treatment group on the basis of a covariate // *Journal of educational and behavioral statistics*. 1977, vol. 2, № 1. P. 1–26. DOI: 10.3102/10769986002001001.

23. *Rubin D. B.* Bayesian inference for causal effects: The role of randomization // *The annals of statistics*. 1978, vol. 6, № 1. P. 34–58. URL: <https://www.jstor.org/stable/2958688> (дата обращения: 05.10.2025).

24. *Pearl J.* Causal diagrams for empirical research // *Biometrika*. 1995, vol. 82, № 4. P. 669–688. DOI: 10.1093/biomet/82.4.669.

25. *Pearl J.* Causality. 2nd ed. Cambridge: Cambridge University Press, 2009. 464 p. ISBN 978-0-521-89560-6. DOI: 10.1017/CBO9780511803161.

26. *Bollen K. A., Pearl J.* Eight myths about causality and structural equation models // *Handbook of causal analysis for social research* / Ed. by S. L. Morgan. Dordrecht: Springer Netherlands, 2013. P. 301–328. DOI: 10.1007/978-94-007-6094-3_15.

27. *Markus K. A.* Causal effects and counterfactual conditionals: Contrasting Rubin, Lewis and Pearl // *Economics & philosophy*. 2021, vol. 37, № 3. P. 441–461. DOI: 10.1017/S0266267120000437.

28. *Imbens G. W.* Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics // *Journal of economic literature*. 2020, vol. 58, № 4. P. 1129–1179. DOI: 10.1257/jel.20191597.

29. *Shpitser I., Pearl J.* Effects of treatment on the treated: Identification and generalization // *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*. Arlington: AUAI Press, 2009. P. 514–521. DOI: 10.5555/1795114.1795174.

30. *Pearl J., Glymour M., Jewell N. P.* Causal inference in statistics: A primer. UK: John Wiley & Sons, 2016. 160 p. ISBN 978-1-119-18686-1.

31. *Cunningham S.* Causal inference: The mixtape. New Haven: Yale University Press, 2021. 328 p. ISBN 978-0-300-25168-5.

32. *Hernán M., Robins J. M.* Causal inference: What if. Boca Raton: Chapman & Hall/CRC, 2020. 350 p. ISBN 978-1-420-07616-5.

33. *Huntington-Klein N.* The effect: An introduction to research design and causality. 2nd ed. Boca Raton: Chapman and Hall/CRC, 2025. 686 p. ISBN 978-1-032-12578-7.

34. Ениколопов П. С. * Оценивание эффектов воздействия // *Квантиль*. 2009, № 6. С. 3–14.

35. Ньюи У. Эффекты воздействия // *Квантиль*. 2009. № 6. С. 15–23.

36. Седашов Е. А. Методы каузального анализа в современной политической науке // *Политическая наука*. 2021, № 1. С. 98–115. DOI: 10.31249/poln/2021.01.04. EDN: SWFGOG.

37. Де Мескута И. Б., Фаулер Э. Статистика без подвоха: методы критического анализа данных и причинного вывода / Пер. с англ. В. С. Яценкова. М.: ДМК Пресс, 2023. 453 с. ISBN 978-5-93700-240-2.

38. *Imbens G. W., Rubin D. B.* Causal inference in statistics, social, and biomedical sciences. Cambridge: Cambridge University Press, 2015. 625 p. ISBN 978-0-521-88588-1. DOI: 10.1017/CBO9781139025751.

39. *Imbens G. W., Xu Y.* LaLonde (1986) after nearly four decades: Lessons learned. arXiv:2406.00827, 2024. DOI: 10.48550/arXiv.2406.00827.

40. *Smith G. C., Pell J. P.* Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials // *BMJ*. 2003, vol. 327, № 7429. P. 1459–1461. DOI: 10.1136/bmj.327.7429.1459.

41. *Yeh R. W., Valsdottir L. R., Yeh M. W., Shen C., Kramer D. B., Strom J. B., Secemsky E. A., Healy J. L., Domeier R. M., Kazi D. S., Nallamothu B. K.* Parachute use to prevent death and major trauma when jumping from aircraft: Randomized controlled trial // *BMJ*. 2018, vol. 363. k5094. DOI: 10.1136/bmj.k5094.

* Внесен в список иностранных агентов Минюста РФ.

42. Rubin D. B. Comment: Which ifs have causal answers // Journal of the American statistical association. 1986, vol. 81, № 396. P. 961–962. DOI: 10.1080/01621459.1986.10478355.

43. Cole S. R., Frangakis C. E. The consistency statement in causal inference: A definition or an assumption? // Epidemiology. 2009, vol. 20, № 1. P. 3–5. DOI: 10.1097/EDE.0b013e3181818ef366.

44. VanderWeele T. J. Concerning the consistency assumption in causal inference // Epidemiology. 2009, vol. 20, № 6. P. 880–883. DOI: 10.1097/ede.0b013e3181818181bd5638.

45. Pearl J. On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? // Epidemiology. 2010, vol. 21, № 6. P. 872–875. DOI: 10.1097/ede.0b013e3181818181f5d3fd.

46. Naimi A. I., Whitcomb B. W. Defining and identifying average treatment effects // American journal of epidemiology. 2023, vol. 192, № 5. P. 685–687. DOI: 10.1093/aje/kwad012.

47. VanderWeele T. J., Ding P. Sensitivity analysis in observational research: Introducing the E-value // Annals of internal medicine. 2017, vol. 167, № 4. P. 268–274. DOI: 10.7326/M16-2607.

48. Cinelli C., Hazlett C. Making sense of sensitivity: Extending omitted variable bias // Journal of the Royal Statistical Society. Series B. 2020, vol. 82, № 1. P. 39–67. DOI: 10.1111/rssb.12348.

49. Rosenbaum P. R. Modern algorithms for matching in observational studies // Annual review of statistics and its application. 2020, vol. 7. P. 143–176. DOI: 10.1146/annurev-statistics-031219-041058.

50. Stolley P. D. When genius errs: R.A. Fisher and the lung cancer controversy // American journal of epidemiology. 1991, vol. 133, № 5. P. 416–425. DOI: 10.1093/oxfordjournals.aje.a115906.

51. Cornfield J., Haenszel W., Hammond E. C., Lilienfeld A. M., Shimkin M. B., Wynder E. L. Smoking and lung cancer: Recent evidence and a discussion of some questions // Journal of the National Cancer Institute. 1959, vol. 22, № 1. P. 173–203. DOI: 10.1093/jnci/22.1.173.

52. Ding P. A first course in causal inference. Chapman and Hall/CRC, 2024. 464 p. ISBN 978-1-032-75862-6.

53. Greifer N., Stuart E. A. Matching methods for confounder adjustment: An addition to the epidemiologist's toolbox // Epidemiologic reviews. 2021, vol. 43, № 1. P. 118–129. DOI: 10.1093/epirev/mxab003.

54. King G., Zeng G. The dangers of extreme counterfactuals // Political analysis. 2006, vol. 14, № 2. P. 131–159. DOI: 10.1093/pan/mpj004.

55. King G., Lucas C., Nielsen R. A. The balance-sample size frontier in matching methods for causal inference // American journal of political science. 2017, vol. 61, № 2. P. 473–489. DOI: 10.1111/ajps.12272.

56. Greifer N. Assessing balance // Cran.R Project: [сайт]. 29.05.2025. URL: <https://cran.r-project.org/web/packages/MatchIt/vignettes/assessing-balance.html> (дата обращения: 05.10.2025).

57. Angrist J. D., Pischke J. S. Mostly harmless econometrics: An empiricist's companion. Princeton: Princeton University Press, 2009. 274 p. ISBN 978-0-691-12035-5. DOI: 10.1515/9781400829828.

58. Angrist J. D., Imbens G. W., Rubin D. B. Identification of causal effects using instrumental variables // Journal of the American statistical association. 1996, vol. 91, № 434. P. 444–455. DOI: 10.2307/2291629.

59. Cattaneo M. D., Idrobo N., Titiunik R. A. Practical introduction to regression discontinuity designs: Foundations. Cambridge: Cambridge University Press, 2020. 120 p. ISBN 978-1-108-71020-6. DOI: 10.1017/9781108684606.

60. Кузьмина Ю. В. Метод разрывной регрессии и метод отбора подбного по вероятности для оценки эффекта одного года обучения: опыт применения на примере данных PISA 2009 // Социология: методология, методы, математическое моделирование. 2014, № 38. С. 7–37. EDN: SZSTOZ.

61. Abadie A. Semiparametric instrumental variable estimation of treatment response models // Journal of econometrics. 2003, vol. 113, № 2. P. 231–263. DOI: 10.1016/S0304-4076(02)00201-4.

62. Angrist J. D., Imbens G. W. Two-stage least squares estimation of average causal effects in models with variable treatment intensity // Journal of the American statistical association. 1995, vol. 90, № 430. P. 431–442. DOI: 10.1080/01621459.1995.10476535.

63. Angrist J. D., Krueger A. B. Empirical strategies in labor economics // Handbook of labor economics. 1999, vol. 3. P. 1277–1366. DOI: 10.1016/S1573-4463(99)03004-7.

64. Cattaneo M. D., Titiunik R. A. Regression discontinuity designs // Annual review of economics. 2022, vol. 14, № 1. P. 821–851. DOI: 10.1146/annurev-economics-051520-021409.

65. Cattaneo M. D., Idrobo N., Titiunik R. A. Practical introduction to regression discontinuity designs: Extensions. Cambridge: Cambridge University Press, 2024. 122 p. ISBN 978-1-009-46232-7. DOI: 10.1017/9781009441896.

66. Freedman B. Equipoise and the ethics of clinical research // The New England journal of medicine. 1987, vol. 317, № 3. P. 141–145. DOI: 10.1056/nejm198707163170304.

67. Liu J., Liu Z., Xu Y. A practical guide to estimating conditional marginal effects: Modern approaches. arXiv, 2025. 2504.01355v1. DOI: 10.48550/arXiv.2504.01355.
68. Wager S., Athey S. Estimation and inference of heterogeneous treatment effects using random forests // Journal of the American statistical association. 2018, vol. 113, № 523. P. 1228–1242. DOI: 10.1080/01621459.2017.1319839.
69. Brand J. E., Zhou X., Xie Y. Recent developments in causal inference and machine learning // Annual review of sociology. 2023, vol. 49. P. 81–110. DOI: 10.1146/annurev-soc-030420-015345.
70. Bansak K., Hainmueller J., Hopkins D. J., Yamamoto T. Conjoint survey experiments // Advances in experimental political science / Eds. by J. N. Druckman, D. P. Green. Cambridge: Cambridge University Press, 2021. P. 19–41. DOI: 10.1017/9781108777919.004.
71. VanderWeele T. J. Explanation in causal inference: Methods for mediation and interaction. New York: Oxford University Press, 2015. 728 p. ISBN 978-0-19-932587-0.
72. Bertrand M., Mullainathan S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination // American economic review. 2004, vol. 94, № 4. P. 991–1013. DOI: 10.1257/0002828042002561.
73. Bessudnov A., Shcherbak A. Ethnic discrimination in multi-ethnic societies: Evidence from Russia // European sociological review. 2020, vol. 36, № 1. P. 104–120. DOI: 10.1093/esr/jcz045.
74. Egami N., Imai K. Causal interaction in factorial experiments: Application to conjoint analysis // Journal of the American statistical association. 2019, vol. 114, № 526. P. 529–540. DOI: 10.1080/01621459.2018.1476246.
75. Imai K., Keele L., Tingley D. A general approach to causal mediation analysis // Psychological methods. 2010, vol. 15, № 4. P. 309–334. DOI: 10.1037/a0020761.
76. Tingley D., Yamamoto T., Hirose K., Keele L., Imai K. Mediation: R package for causal mediation analysis // Journal of statistical software. 2014, vol. 59, № 5. P. 1–38. DOI: 10.18637/jss.v059.i05.
77. Bellani L., Bia M. The long-run effect of childhood poverty and the mediating role of education // Journal of the Royal Statistical Society series A: Statistics in society. 2019, vol. 182, № 1. P. 37–68. DOI: 10.1111/rssa.12388.
78. Daniel R., Zhang J., Farewell D. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets // Biometrical journal. 2021, vol. 63, № 3. P. 528–557. DOI: 10.1002/bimj.201900297.

79. *Morris T. P., Walker A. S., Williamson E. J., White I. R.* Planning a method for covariate adjustment in individually randomised trials: A practical guide // *Trials*. 2022, vol. 23, № 1. 328. DOI: 10.1186/s13063-022-06097-z.

80. *Greenland S.* Noncollapsibility, confounding, and sparse-data bias. Part 1: The oddities of odds // *Journal of clinical epidemiology*. 2021, vol. 138. P. 178–181. DOI: 10.1016/j.jclinepi.2021.06.007.

81. *Greifer N.* Estimating effects after matching // *Cran.R Project*: [сайт]. 29.05.2025. URL: <https://cran.r-project.org/web/packages/MatchIt/vignettes/estimating-effects.html> (дата обращения: 05.10.2025).

82. *Chernozhukov V., Hansen C.* An IV model of quantile treatment effects // *Econometrica*. 2005, vol. 73, № 1. P. 245–261. DOI: 10.1111/j.1468-0262.2005.00570.x.

83. *Abbring J. H., Heckman J. J.* Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation // *Handbook of econometrics*. 2007, vol. 6. P. 5145–5303. DOI: 10.1016/S1573-4412(07)06072-2.

84. *Lu J., Ding P., Dasgupta T.* Treatment effects on ordinal outcomes: Causal estimands and sharp bounds // *Journal of educational and behavioral statistics*. 2018, vol. 43, № 5. P. 540–567. DOI: 10.3102/1076998618776435.

85. *Di Francesco R., Mellace G.* Causal inference for qualitative outcomes. arXiv, 2025. 2502.11691. DOI: 10.48550/arXiv.2502.11691.

86. *Callaway B., Goodman-Bacon A., Sant'Anna P. H.* Difference-in-differences with a continuous treatment. National Bureau of Economic Research, 2024. 59 p. w32117. DOI: 10.3386/w32117.

87. *Imbens G. W.* The role of the propensity score in estimating dose-response functions // *Biometrika*. 2000, vol. 87, № 3. P. 706–710. DOI: 10.1093/biomet/87.3.706.

88. *Imai K., van Dyk D. A.* Causal inference with general treatment regimes: Generalizing the propensity score // *Journal of the American statistical association*. 2004, vol. 99, № 467. P. 854–866. DOI: 10.1198/016214504000001187.

89. *Вулдридж Д. М.* Оценивание методом «разность разностей» // *Квантиль*. 2009, № 6. С. 15–23.

90. *Baker A. C., Larcker D. F., Wang C. C.* How much should we trust staggered difference-in-differences estimates? // *Journal of financial economics*. 2022, vol. 144, № 2. P. 370–395. DOI: 10.1016/j.jfineco.2022.01.004.

91. *de Chaisemartin C., d'Haultfoeulle X.* Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey // *The econometrics journal*. 2023, vol. 26, № 3. P. 1–30. DOI: 10.1093/ectj/utac017.

92. *Roth J., Sant'Anna P. H., Bilinski A., Poe J.* What's trending in difference-in-differences? A synthesis of the recent econometrics literature // *Journal*

of econometrics. 2023, vol. 235, № 2. P. 2218–2244. DOI: 10.1016/j.jeconom.2023.03.008.

93. *Xu Y.* Causal inference with time-series cross-sectional data: A reflection // Oxford handbook of engaged methodological pluralism in political science (online edition) / Eds. By J. M. Box-Steffensmeier, D. P. Christenson, V. Sinclair-Chapman. 2023. 34 p. DOI: 10.1093/oxfordhb/9780192868282.013.30.

94. *Liu L., Wang Y., Xu Y.* A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data // American journal of political science. 2024, vol. 68, № 1. P. 160–176. DOI: 10.1111/ajps.12723.

Сведения об авторе:

Соколов Борис Олегович

Кандидат политических наук; ведущий научный сотрудник
Лаборатории сравнительных социальных исследований
им. Р.Ф. Инглхарта Национального исследовательского
университета «Высшая школа экономики»; доцент
департамента социологии Петербургского кампуса
Национального исследовательского университета
«Высшая школа экономики».

Тел.: +7 (911) 227 59 42.

SPIN-RSCI: 6754-8404.

ResearcherID: L-6331-2015.

Scopus AuthorID: 56415959900.

CAUSAL ESTIMANDS FOR SOCIAL SCIENTIFIC RESEARCH

Sokolov Boris O.,

Ronald F. Inglehart Laboratory for Comparative Social Research, HSE
University; Department of Sociology, HSE Campus in St. Petersburg,
St. Petersburg, Russian Federation
bssokolov@hse.ru
ORCID: 0000-0002-5151-8147

For citation: Sokolov B.O. causal estimands for social scientific research, *Sotsiologiya: 4M (Sociology: methodology, methods, mathematical modeling)*, 2025, no 61, p. 7-76. DOI: 10.19181/4m.2025.34.2.1.

Abstract. This article reviews a set of estimands commonly used in modern scientific and applied research to operationalize causal inquiries within the Rubin Causal Model (RCM). I first introduce the basic average treatment effects (ATE, ATT, ATC) and then describe their main extensions, including local and conditional treatment effects, causal interactions, causal mediation, multivalued or continuous treatments, and longitudinal treatment effects. For each estimand, I provide a substantive explanation along with examples of research questions it can address. For the most popular estimands, I also discuss the key assumptions required for their nonparametric identification and statistical methods available for their estimation.

Keywords: causal inference; treatment effect; Rubin Causal Model; evidence-based policy; estimand; quantitative methods.

Acknowledgments: The article was prepared within the framework of the HSE University Basic Research Program. The author is grateful to the editor and the anonymous reviewer, as well as to Alexey Bessudnov, Yegor Lazarev, Elizaveta Popkova, Maxim Rudnev and Yuri Rykov, for their valuable recommendations and comments.

References

1. Enikolopov R.S. Evidence based development economics: Nobel Prize in Economic Sciences 2019 (in Russian), *Voprosy Ekonomiki*, 2020, no. 1, p. 5–17. DOI: 10.32609/0042-8736-2020-1-5-17.
2. Serebrennikov D.E., Kuzmina Yu.V. Field Experiments and the Rubin Causal Model: Review of Approaches and Current Research (in Russian), *Economic Sociology*, 2021, vol. 22, no. 4, p. 117–139. DOI: 10.17323/1726-3247-2021-4-117-139.
3. Heckman J.J. Econometric causality, *International statistical review*, 2008, vol. 76, no. 1, p. 1–27. DOI: 10.1111/j.1751-5823.2007.00024.x.
4. Sonin K.I. Instead of laboratory: Analyzing data from natural experiments (Nobel Memorial Prize in Economic Sciences 2021) (in Russian), *Voprosy Ekonomiki*, 2022, no. 1, p. 5–22. DOI: 10.32609/0042-8736-2022-1-5-22.
5. Hitchcock C. “Probabilistic causation”, in: *The Stanford encyclopedia of philosophy (Spring 2021 Edition)*, ed. by E. N. Zalta. 2021. URL: <https://plato.stanford.edu/archives/spr2021/entries/causation-probabilistic/> (date of access: 05.10.2025).
6. Imbens G.W. Causal inference in the social sciences, *Annual review of statistics and its application*, 2024, vol. 11, p. 123–152. DOI: 10.1146/annurev-statistics-033121-114601.
7. Hassell H.J., Holbein J.B. Navigating potential pitfalls in difference-in-differences designs: Reconciling conflicting findings on mass shootings' effect on electoral outcomes, *American political science review*, 2024, vol. 119, no. 1, p. 240–260. DOI: 10.1017/S0003055424000108.
8. Greifer N., Stuart E.A. *Choosing the estimand when matching or weighting in observational studies*. arXiv:2106.10577 [Stat], June 2021. DOI: 10.48550/arXiv.2106.10577.
9. Heiss A. “Demystifying causal inference estimands: ATE, ATT, and ATU”, in: *Andrew Heiss [site]*. 21.03.2024. DOI: 10.59350/c9z3a-rcq16. URL: <https://www.andrewheiss.com/blog/2024/03/21/demystifying-ate-att-atu/> (date of access: 05.10.2025).
10. Barrett M., D'Agostino McGowan L., Gerke T. “Causal inference in R”, in: *R-Causal [site]*. 21.08.2025. URL: <https://www.r-causal.org/> (date of access: 05.10.2025).

11. Lundberg I., Johnson R., Stewart B.M. What is your estimand? Defining the target quantity connects statistical evidence to theory, *American sociological review*, 2021, vol. 86, no. 3, p. 532–565. DOI: 10.1177/00031224211004187.
12. Neyman (Splawa-Neyman) J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9, *Roczniki Nauk Rolniczych*, tom X. 1923 [in Polish]; reprinted in: *Statistical science*, 1990, vol. 5, no. 4, p. 465–472 / Transl. by D. M. Dabrowska, T. P. Speed. URL: <https://www.jstor.org/stable/2245382> (date of access: 05.10.2025).
13. Holland P.W. Statistics and causal inference, *Journal of the American statistical association*, 1986, vol. 81, no. 396, p. 945–960. DOI: 10.1080/01621459.1986.10478354.
14. Rubin D.B. [On the application of probability theory to agricultural experiments. Essay on principles. Section 9.] Comment: Neyman (1923) and causal inference in experiments and observational studies, *Statistical science*, 1990, vol. 5, no. 4, p. 472–480. URL: <https://www.jstor.org/stable/2245383> (date of access: 05.10.2025).
15. Angrist J.D., Imbens G.W., Rubin D.B. Identification of causal effects using instrumental variables: Rejoinder, *Journal of the American statistical association*, 1996, vol. 91, no. 434, p. 468–472. DOI: 10.2307/2291634.
16. Rubin D.B. Causal inference using potential outcomes: Design, modeling, decisions, *Journal of the American statistical association*, 2005, vol. 100, no. 469, p. 322–331. DOI: 10.1198/016214504000001880.
17. Heckman J.J. Identification of causal effects using instrumental variables: Comment, *Journal of the American statistical association*, 1996, vol. 91, no. 434, p. 459–462. DOI: 10.2307/2291631.
18. Pearl J. Causal inference: History, perspectives, adventures, and unification (An interview with Judea Pearl), *Observational studies*, 2022, vol. 8, no. 2, p. 23–36. DOI: 10.1353/obs.2022.0007.
19. Rubin D.B. Matching to remove bias in observational studies, *Biometrics*, 1973, vol. 29, no. 1, p. 159–183. DOI: 10.2307/2529684.
20. Rubin D.B. The use of matched sampling and regression adjustment to remove bias in observational studies, *Biometrics*, 1973, vol. 29, no. 1, p. 185–203. DOI: 10.2307/2529685.

21. Rubin D.B. Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of educational psychology*, 1974, vol. 66, no. 5, p. 688–701. DOI: 10.1037/h0037350.
22. Rubin D.B. Assignment to treatment group on the basis of a covariate, *Journal of educational and behavioral statistics*, 1977, vol. 2, no. 1, p. 1–26. DOI: 10.3102/10769986002001001.
23. Rubin D.B. Bayesian inference for causal effects: The role of randomization, *The annals of statistics*, 1978, vol. 6, no. 1, p. 34–58. URL: <https://www.jstor.org/stable/2958688> (date of access: 05.10.2025).
24. Pearl J. Causal diagrams for empirical research, *Biometrika*, 1995, vol. 82, no. 4, p. 669–688. DOI: 10.1093/biomet/82.4.669.
25. Pearl J. *Causality*. 2nd ed. Cambridge: Cambridge University Press, 2009. 464 p. ISBN 978-0-521-89560-6. DOI: 10.1017/CBO9780511803161.
26. Bollen K.A., Pearl J. “Eight myths about causality and structural equation models”, in: *Handbook of causal analysis for social research*, ed. by S. L. Morgan. Dordrecht: Springer Netherlands, 2013. P. 301–328. DOI: 10.1007/978-94-007-6094-3_15.
27. Markus K.A. Causal effects and counterfactual conditionals: Contrasting Rubin, Lewis and Pearl, *Economics & philosophy*, 2021, vol. 37, no. 3, p. 441–461. DOI: 10.1017/S0266267120000437.
28. Imbens G.W. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics, *Journal of economic literature*, 2020, vol. 58, no. 4, p. 1129–1179. DOI: 10.1257/jel.20191597.
29. Shpitser I., Pearl J. “Effects of treatment on the treated: Identification and generalization”, in: *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*. Arlington: AUAI Press, 2009. P. 514–521. DOI: 10.5555/1795114.1795174.
30. Pearl J., Glymour M., Jewell N.P. *Causal inference in statistics: A primer*. UK: John Wiley & Sons, 2016. 160 p. ISBN 978-1-119-18686-1.
31. Cunningham S. *Causal inference: The mixtape*. New Haven: Yale University Press, 2021. 328 p. ISBN 978-0-300-25168-5.
32. Hernán M., Robins J.M. *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC, 2020. 350 p. ISBN 978-1-420-07616-5.

33. Huntington-Klein N. *The effect: An introduction to research design and causality*. 2nd ed. Boca Raton: Chapman and Hall/CRC, 2025. 686 p. ISBN 978-1-032-12578-7.
34. Enikolopov R.S. Estimation of treatment effects (in Russian), *Quantile*, 2009, no. 6, p. 3–14.
35. Newey W. Treatment effects (in Russian), *Quantile*, 2009, no. 6, p. 15–23.
36. Sedashov E.A. Methods of causal analysis in modern political science (in Russian), *Politicheskaya Nauka (Political Science)*, 2021, no. 1, p. 98–115. DOI: 10.31249/poln/2021.01.04.
37. De Mesquita E.B., Fowler A. Statistics without a trick: Methods of critical data analysis and causal inference (transl., in Russian). Moscow: DMK Press, 2023. 453 p. ISBN 978-5-93700-240-2.
38. Imbens G.W., Rubin D.B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press, 2015. 625 p. ISBN 978-0-521-88588-1. DOI: 10.1017/CBO9781139025751.
39. Imbens G.W., Xu Y. *LaLonde (1986) after nearly four decades: Lessons learned*. arXiv:2406.00827, 2024. DOI: 10.48550/arXiv.2406.00827.
40. Smith G.C., Pell J.P. Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials, *BMJ*, 2003, vol. 327, no. 7429, p. 1459–1461. DOI: 10.1136/bmj.327.7429.1459.
41. Yeh R.W., Valsdottir L.R., Yeh M.W., Shen C., Kramer D.B., Strom J.B., Secemsky E.A., Healy J.L., Domeier R.M., Kazi D.S., Nallamothu B.K. Parachute use to prevent death and major trauma when jumping from aircraft: Randomized controlled trial, *BMJ*, 2018, vol. 363. k5094. DOI: 10.1136/bmj.k5094.
42. Rubin D.B. Comment: Which ifs have causal answers, *Journal of the American statistical association*, 1986, vol. 81, no. 396, p. 961–962. DOI: 10.1080/01621459.1986.10478355.
43. Cole S.R., Frangakis C.E. The consistency statement in causal inference: A definition or an assumption? *Epidemiology*, 2009, vol. 20, no. 1, p. 3–5. DOI: 10.1097/EDE.0b013e31818ef366.
44. VanderWeele T.J. Concerning the consistency assumption in causal inference, *Epidemiology*, 2009, vol. 20, no. 6, p. 880–883. DOI: 10.1097/ede.0b013e3181bd5638.

45. Pearl J. On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? *Epidemiology*, 2010, vol. 21, no. 6, p. 872–875. DOI: 10.1097/ede.0b013e3181f5d3fd.
46. Naimi A.I., Whitcomb B.W. Defining and identifying average treatment effects, *American journal of epidemiology*, 2023, vol. 192, no. 5, p. 685–687. DOI: 10.1093/aje/kwad012.
47. VanderWeele T.J., Ding P. Sensitivity analysis in observational research: Introducing the E-value, *Annals of internal medicine*, 2017, vol. 167, no. 4, p. 268–274. DOI: 10.7326/M16-2607.
48. Cinelli C., Hazlett C. Making sense of sensitivity: Extending omitted variable bias, *Journal of the Royal Statistical Society. Series B*, 2020, vol. 82, no. 1, p. 39–67. DOI: 10.1111/rssb.12348.
49. Rosenbaum P.R. Modern algorithms for matching in observational studies, *Annual review of statistics and its application*, 2020, vol. 7, p. 143–176. DOI: 10.1146/annurev-statistics-031219-041058.
50. Stolley P.D. When genius errs: R.A. Fisher and the lung cancer controversy, *American journal of epidemiology*, 1991, vol. 133, no. 5, p. 416–425. DOI: 10.1093/oxfordjournals.aje.a115906.
51. Cornfield J., Haenszel W., Hammond E.C., Lilienfeld A.M., Shimkin M.B., Wynder E.L. Smoking and lung cancer: Recent evidence and a discussion of some questions, *Journal of the National Cancer Institute*, 1959, vol. 22, no. 1, p. 173–203. DOI: 10.1093/jnci/22.1.173.
52. Ding P. *A first course in causal inference*. Chapman and Hall/CRC, 2024. 464 p. ISBN 978-1-032-75862-6.
53. Greifer N., Stuart E.A. Matching methods for confounder adjustment: An addition to the epidemiologist's toolbox, *Epidemiologic reviews*, 2021, vol. 43, no. 1, p. 118–129. DOI: 10.1093/epirev/mxab003.
54. King G., Zeng L. The dangers of extreme counterfactuals, *Political analysis*, 2006, vol. 14, no. 2, p. 131–159. DOI: 10.1093/pan/mpj004.
55. King G., Lucas C., Nielsen R.A. The balance-sample size frontier in matching methods for causal inference, *American journal of political science*, 2017, vol. 61, no. 2, p. 473–489. DOI: 10.1111/ajps.12272.
56. Greifer N. “Assessing balance”, in: *Cran.R Project*: [site]. 29.05.2025. URL: <https://cran.r-project.org/web/packages/MatchIt/vignettes/assessing-balance.html> (date of access: 05.10.2025).

57. Angrist J.D., Pischke J.S. *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press, 2009. 274 p. ISBN 978-0-691-12035-5. DOI: 10.1515/9781400829828.
58. Angrist J.D., Imbens G.W., Rubin D.B. Identification of causal effects using instrumental variables, *Journal of the American statistical association*, 1996, vol. 91, no. 434, p. 444–455. DOI: 10.2307/2291629.
59. Cattaneo M.D., Idrobo N., Titiunik R.A. *Practical introduction to regression discontinuity designs: Foundations*. Cambridge: Cambridge University Press, 2020. 120 p. ISBN 978-1-108-71020-6. DOI: 10.1017/9781108684606.
60. Kuzmina Yu. V. Regression discontinuity design method and propensity score matching (PSM) in one year of schooling effect evaluation: an experience of applying to PISA 2009 data (in Russian), *Sociology: methodology, methods, mathematical modeling (Sociology: 4M)*, 2014, no. 38, p. 7–37.
61. Abadie A. Semiparametric instrumental variable estimation of treatment response models, *Journal of econometrics*, 2003, vol. 113, no. 2, p. 231–263. DOI: 10.1016/S0304-4076(02)00201-4.
62. Angrist J.D., Imbens G.W. Two-stage least squares estimation of average causal effects in models with variable treatment intensity, *Journal of the American statistical association*, 1995, vol. 90, no. 430, p. 431–442. DOI: 10.1080/01621459.1995.10476535.
63. Angrist J.D., Krueger A.B. “Empirical strategies in labor economics”, in: *Handbook of labor economics*, 1999, vol. 3, p. 1277–1366. DOI: 10.1016/S1573-4463(99)03004-7.
64. Cattaneo M.D., Titiunik R.A. Regression discontinuity designs, *Annual review of economics*, 2022, vol. 14, no. 1, p. 821–851. DOI: 10.1146/annurev-economics-051520-021409.
65. Cattaneo M.D., Idrobo N., Titiunik R.A. *Practical introduction to regression discontinuity designs: Extensions*. Cambridge: Cambridge University Press, 2024. 122 p. ISBN 978-1-009-46232-7. DOI: 10.1017/9781009441896.
66. Freedman B. Equipoise and the ethics of clinical research, *The New England journal of medicine*, 1987, vol. 317, no. 3, p. 141–145. DOI: 10.1056/nejm198707163170304.

67. Liu J., Liu Z., Xu Y. *A practical guide to estimating conditional marginal effects: Modern approaches*. arXiv, 2025. 2504.01355v1. DOI: 10.48550/arXiv.2504.01355.
68. Wager S., Athey S. Estimation and inference of heterogeneous treatment effects using random forests, *Journal of the American statistical association*, 2018, vol. 113, no. 523, p. 1228–1242. DOI: 10.1080/01621459.2017.1319839.
69. Brand J.E., Zhou X., Xie Y. Recent developments in causal inference and machine learning, *Annual review of sociology*, 2023, vol. 49, p. 81–110. DOI: 10.1146/annurev-soc-030420-015345.
70. Bansak K., Hainmueller J., Hopkins D.J., Yamamoto T. “Conjoint survey experiments”, in: *Advances in experimental political science*, ed. by J. N. Druckman, D. P. Green. Cambridge: Cambridge University Press, 2021. P. 19–41. DOI: 10.1017/9781108777919.004.
71. VanderWeele T.J. *Explanation in causal inference: Methods for mediation and interaction*. New York: Oxford University Press, 2015. 728 p. ISBN 978-0-19-932587-0.
72. Bertrand M., Mullainathan S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination, *American economic review*, 2004, vol. 94, no. 4, p. 991–1013. DOI: 10.1257/0002828042002561.
73. Bessudnov A., Shcherbak A. Ethnic discrimination in multi-ethnic societies: Evidence from Russia, *European sociological review*, 2020, vol. 36, no. 1, p. 104–120. DOI: 10.1093/esr/jcz045.
74. Egami N., Imai K. Causal interaction in factorial experiments: Application to conjoint analysis, *Journal of the American statistical association*, 2019, vol. 114, no. 526, p. 529–540. DOI: 10.1080/01621459.2018.1476246.
75. Imai K., Keele L., Tingley D. A general approach to causal mediation analysis, *Psychological methods*, 2010, vol. 15, no. 4, p. 309–334. DOI: 10.1037/a0020761.
76. Tingley D., Yamamoto T., Hirose K., Keele L., Imai K. Mediation: R package for causal mediation analysis, *Journal of statistical software*, 2014, vol. 59, no. 5, p. 1–38. DOI: 10.18637/jss.v059.i05.
77. Bellani L., Bia M. The long-run effect of childhood poverty and the mediating role of education, *Journal of the Royal Statistical*

- Society series A: Statistics in society*, 2019, vol. 182, no. 1, p. 37–68. DOI: 10.1111/rssa.12388.
78. Daniel R., Zhang J., Farewell D. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets, *Biometrical journal*, 2021, vol. 63, no. 3, p. 528–557. DOI: 10.1002/bimj.201900297.
79. Morris T.P., Walker A.S., Williamson E.J., White I.R. Planning a method for covariate adjustment in individually randomised trials: A practical guide, *Trials*, 2022, vol. 23, no. 1, 328. DOI: 10.1186/s13063-022-06097-z.
80. Greenland S. Noncollapsibility, confounding, and sparse-data bias. Part 1: The oddities of odds, *Journal of clinical epidemiology*, 2021, vol. 138, p. 178–181. DOI: 10.1016/j.jclinepi.2021.06.007.
81. Greifer N. “Estimating effects after matching”, in: *Cran.R Project* [site]. 29.05.2025. URL: <https://cran.r-project.org/web/packages/MatchIt/vignettes/estimating-effects.html> (date of access: 05.10.2025).
82. Chernozhukov V., Hansen C. An IV model of quantile treatment effects, *Econometrica*, 2005, vol. 73, no. 1, p. 245–261. DOI: 10.1111/j.1468-0262.2005.00570.x.
83. Abbring J.H., Heckman J.J. “Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation”, in: *Handbook of econometrics*, 2007, vol. 6, p. 5145–5303. DOI: 10.1016/S1573-4412(07)06072-2.
84. Lu J., Ding P., Dasgupta T. Treatment effects on ordinal outcomes: Causal estimands and sharp bounds, *Journal of educational and behavioral statistics*, 2018, vol. 43, no. 5, p. 540–567. DOI: 10.3102/1076998618776435.
85. Di Francesco R., Mellace G. *Causal inference for qualitative outcomes*. arXiv, 2025. 2502.11691. DOI: 10.48550/arXiv.2502.11691.
86. Callaway B., Goodman-Bacon A., Sant'Anna P.H. *Difference-in-differences with a continuous treatment*. National Bureau of Economic Research, 2024. 59 p. w32117. DOI: 10.3386/w32117.
87. Imbens G.W. The role of the propensity score in estimating dose-response functions, *Biometrika*, 2000, vol. 87, no. 3, p. 706–710. DOI: 10.1093/biomet/87.3.706.

88. Imai K., van Dyk D.A. Causal inference with general treatment regimes: Generalizing the propensity score, *Journal of the American statistical association*, 2004, vol. 99, no. 467, p. 854–866. DOI: 10.1198/016214504000001187.
89. Wooldridge J. M. Difference-in-differences estimation (in Russian), *Quantile*, 2009, no. 6, p. 15–23.
90. Baker A.C., Larcker D.F., Wang C.C. How much should we trust staggered difference-in-differences estimates? *Journal of financial economics*, 2022, vol. 144, no. 2, p. 370–395. DOI: 10.1016/j.jfineco.2022.01.004.
91. De Chaisemartin C., d'Haultfoeuille X. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey, *The econometrics journal*, 2023, vol. 26, no. 3, p. 1–30. DOI: 10.1093/ectj/utac017.
92. Roth J., Sant'Anna P.H., Bilinski A., Poe J. What's trending in difference-in-differences? A synthesis of the recent econometrics literature, *Journal of econometrics*, 2023, vol. 235, no. 2, p. 2218–2244. DOI: 10.1016/j.jeconom.2023.03.008.
93. Xu Y. “Causal inference with time-series cross-sectional data: A reflection”, in: *Oxford handbook of engaged methodological pluralism in political science (online edition)*, ed. by J. M. Box-Steffensmeier, D. P. Christenson, V. Sinclair-Chapman. 2023. 34 p. DOI: 10.1093/oxfordhb/9780192868282.013.30.
94. Liu L., Wang Y., Xu Y. A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data, *American journal of political science*, 2024, vol. 68, no. 1, p. 160–176. DOI: 10.1111/ajps.12723.

Information about the author

Boris O. Sokolov

Candidate of Science in Political Science; Leading Research Fellow,
Ronald F. Inglehart Laboratory for Comparative Social Research,
HSE University; Associate Professor, Department of Sociology,
HSE Campus in St. Petersburg

SPIN-RSCI: 6754-8404

ResearcherID: L-6331-2015

Scopus AuthorID: 56415959900