Transportation Research Part A 192 (2025) 104324

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Causal inference for transport research

Daniel J. Graham

Centre for Transport Engineering & Modelling, Imperial College London, London, SW7 2AZ, UK

ARTICLE INFO

Keywords:
Causality
Identification
Estimation
Intervention
Treatment
Potential outcome

ABSTRACT

This paper provides a consolidated overview of the statistical literature on causal inference, emphasising its relevance and applicability for transportation research. It outlines a framework for causal identification based on the concept of potential outcomes and provides a summary of core contemporary methods that can be used for estimation. Typical challenges encountered in identifying cause–effect relationships in applied transportation research are analysed via case study simulations, and R code to execute and adapt causal estimators is made available. Causal inference can be used to obtain unbiased and consistent estimates of causal effects in non-experimental settings when interventions or exposures are non-randomly assigned. The paper argues that empirical analyses in transport research are typically conducted in this setting, and consequently, that causal inference has immediate and valuable applicability.

1. Introduction

Transport policy is fundamentally concerned with effecting change. 'Interventions' are made in the hope that they will solve real societal problems (e.g. improve road safety, improve travel times, or alleviate congestion) or change behaviour in some targeted way (e.g. reduce gasoline consumption, decrease driving speeds, induce mode shift).

The rationale for such policy interventions typically rests on assumed cause–effect relationships. It is hoped that by manipulating the system in some way a change in outcomes will be effected (or caused). Such cause–effect assumptions are often sourced from research on transport systems and their wider contexts. Indeed, as an applied discipline, transportation research has a long and extensive track record of direct influence on public policy making. It is used worldwide both *ex-post*, to evaluate historic policy interventions, and *ex-ante*, to inform the design of future interventions .

It is undoubtedly important that cause–effect relationships are understood and quantified for the effective planning and design of transportation policies. However, while a causal interpretation of research is often adopted, the concept of *causality* frequently fails to feature in the conceptual and methodological designs used in transportation research itself. The absence of causal reasoning, particularly in data collation and model specification, can lead to ambiguous, spurious, and even downright misleading inferences. All too often in transport research, correlation is mistaken for causality in interpretation of findings.

There is now a well established literature on *causal inference* that is used routinely across the physical, biological and social sciences to develop the conditions and methods required to draw causal conclusion from statistical analyses. This is based on the general principle that we focus our attention on identifying the *causal effects* of *interventions* (or *treatments*) on *outcomes* of interest, while adjusting for the extraneous influences (or *biases*) that can arise from other aspects of the process that we believe generated the data.

The aim of this paper is to provide a comprehensive summary of the statistical literature on causal inference for use by transport researchers. The paper is not a review of existing applications of causal inference within transport research. Rather, it provides a survey of the core concepts and methods that form the statistical field of causal inference, with insights on how these can be

E-mail address: d.j.graham@imperial.ac.uk.

https://doi.org/10.1016/j.tra.2024.104324
Received 17 November 2023; Received in revised form 24 October 2024; Accepted 11 November 2024

Available online 27 November 2024

fruitfully adopted in empirical work in transport. Causal inference methodologies have yet to feature as strongly in transport as they do in other scientific fields, such as medicine, biostatistics, economics, epidemiology, computer science and the like. Yet, the fundamental empirical challenges that we face in transport are directly comparable to those encountered in these other fields, and consequently, causal inference has immediate and valuable applicability (for a shorter non-technical discussion see Graham, 2021).

The paper is structured as follows. In section two we set notation and define preliminary concepts. The potential outcomes approach to causal inference is introduced in section three, and followed in section four by derivation of the conditions necessary for causal identification. Core methods for estimation of causal quantities are described in sections five and six, under ignorability and non-ignorability respectively. Section seven then illustrates applications of causal methods in transport research via simulation. R code for the execution and generation of these models is available for readers to run and adapt. Conclusions are drawn in the final section.

2. Preliminary concepts

2.1. Notation and structure of causal inference

We first introduce the notation and structure of causal inference and link it to typical concerns in transport research.

We will perform inference in relation to a random vector Z, with realised, or observed data, z, available for units i, $i = 1, \dots, n$, in a sample of n units from a population of interest. As usual, the underlying aim of inference will be to infer properties of the parameters of the joint distribution of Z, $f_Z(z)$, which is potentially unknown and for which we postulate a statistical model. If we posit a parametric model, we will restrict the analytic form of $f_Z(z)$ to a suitable family and assume that it is determined by a finite number of real unknown parameters, $\theta = (\theta^1, \dots, \theta^d)$, that lie in parameter space Ω_θ. The class of densities for finite-dimensional parametric models can be written

$$\mathcal{P} = \left\{ f_Z(z;\theta), z \in \mathcal{Z}, \theta \in \Omega_\theta \subseteq \mathbb{R}^d \right\},$$

where $\mathcal{Z} = \{z : f_Z(z;\theta) > 0\}$ is the sample space in which the data lie and $f_Z(z;\theta)$ is the model function.

Alternatively, we can apply nonparametric (or distribution free) inference and seek to model the data in the absence of a parametric representation. In causal inference we can interchange freely between parametric and nonparametric representations, typically setting up identification in a non parametric fashion and proposing parametric models for estimation.

Causal inference centres around three components of Z: $Z = (Y, D, X)$

1. The *Outcome*, or *Response*, of interest, which we denote Y.
2. The *Treatment* (or *Intervention*) to be studied, which we denote D.
3. The *Covariates*, or measured pre-treatment characteristics, which we denote $X = (X_1, \dots, X_k)$.

We are interested in moments of the conditional density $y = f_{Y|D,X}(d, x; \theta)$. If our aim was simply *prediction*, then emphasis would be on maximising the fit of $\hat{y} = f_{Y|D,X}(d, x; \hat{\theta})$, via some criterion, e.g. $\min n^{-1} \sum_i (y_i - \hat{y}_i)$, but not on $\hat{\theta} = (\hat{\theta}^1, \dots, \hat{\theta}^d)$ per se. Alternatively, if our aim was *associational inference*, then emphasis would be on quantifying associations between (D, X) and Y via an optimal estimator for $\hat{\theta} = (\hat{\theta}^1, \dots, \hat{\theta}^d)$. With *causal inference*, our aim is not just to achieved optimality in estimation, but to *identify* and measure the net *causal* effect of treatment (or intervention) D on outcome Y, that is $\partial Y / \partial D$ or $\Delta D \rightarrow \Delta Y$.

2.1.1. Statistical causality

To motivate our discussion we introduce four case study applications that might represent typical causal concerns in transport policy and analysis. These are summarised below.

Case study	intervention exposure	cause → effect	units
CS1 Road safety	speed cameras	speed cameras → road safety	links
CS2 Travel Behaviour	urban sprawl	urban sprawl → car usage	households
CS3 Network performance	signalling upgrade	signal upgrade → journey times	metro lines
CS4 Environment	low emissions zones	LEZ → air quality	zones

In each case study, we are interested in the causal effect of an intervention (or exposure), D, on an outcome, Y. We analyse this relationship using the data available for our n units. For each case study, random vector $z = (y, d, x)$ could comprise the following data.

	y	d	x
CS1	collision counts	speed camera presence	link characteristics
CS2	household car miles	zone density	household/zone characteristics
CS3	O-D journey times	signalling type	metro line characteristics
CS4	emissions	LEZ designation	zone characteristics

Understanding the causal relationships that characterise our case studies is important for decisions on policy and investment. But what do we actually mean by *causation*, and how does this differ from the familiar statistical concept of *dependence*.

In probabilistic terminology we say that two variables are *associated* if there is an absence of *independence* between them. In the statistical literature the symbol $\perp\!\!\!\perp$ is used to denote independence. If Y and D are independent, that is,

$$Y \perp\!\!\!\perp D,$$

then for every $d \in \mathbb{R}$ and $y \in \mathbb{R}$ their joint density factorises as $f_{D,Y}(d, y) = f_D(d)f_Y(y)$ and their joint expectation as $\mathbb{E}(D \cdot Y) = \mathbb{E}(D) \cdot \mathbb{E}(Y)$. Independence means that knowledge that $D = d$ does not provide any additional information about the distribution of Y.

In contrast, if Y and D are associated, e.g. $Y \not\perp\!\!\!\perp D$, then observing D helps in predicting Y (and vice versa), and certain joint events, captured by the density $f_{D,Y}(d, y)$, tend to occur more often than expected under independence.

Association thus implies statistical dependence between D and Y, but it does not amount to causation. It is, in fact, a necessary, but not sufficient, condition for causation. Causation is a stronger claim. It implies that D has a direct influence on, or *causes*, Y. That is, *manipulating* or *changing* D will cause a change in Y, and make event $Y = y$ more likely. For example, for our case studies, the causality implied by $\Delta D \longrightarrow \Delta Y$ could be that:

- **CS1:** speed camera presence alters the collision rate.
- **CS2:** household location determines car use.
- **CS3:** metro signalling technologies influence journey times.
- **CS4:** low emission zone designation changes air quality.

2.2. Treatment and assignment

The key question that causal inference can help to answer is as follows: what *effect(s)* do *intervention(s)* actually *cause*? Accordingly, our focus will be on the causal effect that a intervention (or some set of interventions) has on an outcome. We may wish to know what the outcome would have been had the intervention not been applied, or if some different intervention been applied.

The emphasis on intervention is key to the underlying the philosophy of causal inference. Holland (1986) distinguishes between *associational inference*, which focuses first on outcomes and attempts to infer the cause of an effect; and *causal inference*, which views the observed outcome as one potential realisation achieved through *manipulation* of an effect. *Intervention* is therefore centre stage in causal inference because we seek to identify the 'effects of causes' rather than deduce the cause of a given effect.

To this end, we will view treatment D as a *random variable* whose manipulation can produce different outcomes, e.g Y. We refer to D as a 'treatment', defined in the broadest sense to encompass any 'regime' which can be manipulated to produce an effect. In fact a treatment can also be an *exposure*, that is, some characteristic that units in the sample experience, rather than have prescribed to them via explicit intervention per se. Treatments are sometimes described as 'natural experiments', so called because, as with an experiment, we can partition a sample of units into sub-populations classified by treatment status (e.g. treated and control for a binary intervention).

In our case studies, treatment D involves: the presence (or absence) of a speed camera e.g. $D \in \{0, 1\}$, the residential density of zones e.g. $D \subseteq \mathbb{R}$ or $D \equiv (d_0, d_1, \ldots, d_m)$, the presence of upgraded signalling e.g. $D \in \{0, 1\}$ or $D \equiv (d_0, d_1, \ldots, d_m)$, and zone designation e.g. $D \in \{0, 1\}$. Thus, our treatment variable can be binary, multivalued, or continuous.

Under true experimental conditions we would have control over the *assignment of treatment*, and could thus actively influence the process through which units receive treatment. For instance, we could assign treatment randomly, and observe differences in outcomes by treatment status. Such an experiment is know as a *randomised control trial (RCT)*, and as we will see later, that randomisation allows causal effects to be identified simply and directly.

In practice, however, active manipulation of treatment assignment is not possible, or desirable, for a variety of reasons; including practical/operational feasibility, safety, ethics, and costs. This challenge often represent a seemingly insurmountable problem for empirical research on cause–effect relations in transport systems. We typically cannot experiment freely with transport policy; or with interventions that affects such important outcomes as economic and environmental welling, road safety, travel behaviour, network performance, and so on.

The causal inference approaches we discuss in this paper aim to quantify effects that occur due to intervention (or exposure) in non-experimental settings, typically for non-randomly assigned treatments. This empirical context is referred to as an *observational setting*, as opposed to experimental, in which data z are generated, not from an experiment, but passively via sampling from a population of units, for example through surveys and measurement.

A central motivation for this paper is the belief that causal questions in transport typically have to be addressed in observational settings. As we will see, causal inference in the observational setting is also challenging. Typically, the processes that generate the data are not only uncontrolled but also largely unknown. This is important, as it is in fact aspects of the so called *Data Generating Process (DGP)*, rather than mere associational inference about the parameters of $f_Z(z; \theta)$, that causal inference seeks to infer.

2.3. Data generating process and identification

An example of a DGP typically encountered in causal inference problems is given below.

$$\begin{aligned} Y|D,X &\sim \mathcal{N}(\beta_0 + \tau D + \beta_1 X, \sigma_Y^2) \\ D|X &\sim \mathcal{B}(\text{expit}(\alpha_0 + \alpha_1 X)) \\ X &\sim \mathcal{N}(\mu_X, \sigma_X^2). \end{aligned}$$

Under this DGP, the outcome Y is determined by a Bernoulli distributed ($\mathcal{B}$) binary treatment D, and by normally distributed ($\mathcal{N}$) covariates X; while the treatment itself is assigned non-randomly as it has dependence on X. Note that expit is the inverse of the logit function, e.g. $\text{expit}(x) = \exp(x)/(1+\exp(x))$.

This DGP essentially captures the relationships assumed for the binary treatments in **CS1**, **CS3** and **CS4**. **CS2** has a continuous treatment, urban density, $D_i \subseteq \mathbb{R}$, so, for example, we could replace the second line above with

$$D_i|X_i \sim \mathcal{N}(\alpha_0 + \alpha_1 X_i, \sigma_{D_i}^2).$$

The DGP then, simply describes the relationships between the three key components of causal inference problems: response (Y), treatment (D), and covariates (X). The DGP is assumed unknown to the analyst, and it is typically not the objective of causal inference to fully recover it, but rather to infer some aspect of it. For example, the parameter $\tau = \partial Y/\partial D$, which describes the causal effect of D on Y, is a common target of inference.

In seeking to infer aspects of the DGP, a fundamental guiding principle of causal inference, addressed prior to and separate from estimation, is the need to achieve *identification*. The concept of identification we draw on here is the standard statistical one.

Definition 1 (*Identification*)**.** *a parameter θ for a family of distributions $\{f_Z(z;\theta) : \theta \in \Omega_\theta\}$ is identifiable if distinct values of θ correspond to distinct pdfs or pmfs. That is, if $\theta \neq \theta'$, then $f_Z(z;\theta)$ is not the same function of z as $f_Z(z;\theta')$*

Definition 2 (*Observational Equivalence*)**.** *if $f_Z(z;\theta) = f_Z(z;\theta')$ then these two structures of the model are said to be observationally equivalent*

The key implication of identification is that if a model is identified then it is logically possible to *uniquely* determine the value of model parameters given an infinite number of observations. In other words, we can derive a unique solution based on observable quantities. The existence of multiple observationally equivalent structures implies that a model is not identified, because if observations from two distributions look identical, we cannot know whether the true value of the parameter is θ or θ' (e.g. not unique).

It is worth noting that the aim of identification contrasts with that of prediction. Consider the model $y = f_X(x;\theta)$. For identification we aim to quantify uniquely some, but perhaps not all, parameters of our model. For example, we may be interested in calculating a single parameter $\hat{\theta}_k = \partial y/\partial x_k$. For prediction, the quantity of interest is $\hat{y}$, a function of x and $\widehat{\theta}$. The aims of prediction and identification are closely related, but not the same. A perfectly identified parameter may, or may not, be particularly helpful in predicting y. Accurate prediction of y however, that is, with relatively low mean squared error, could be achieved via a non-identified model.

The concept of identification is fundamental to causal inference because ultimately we are seeking to quantify a unique set of causal parameters that capture relationships embedded in the DGP.

3. The potential outcomes framework for causal inference

We have noted above that our focus is on interventions, and this is important as a route to identification. We now introduce the *potential outcomes framework* for causal inference, which defines the conditions required for causal identification of treatment effects. This framework was first put forward for binary treatments in a series of papers in the 1970s by Rubin Rubin (e.g. 1973a,b, 1974, 1977, 1978), although Rubin acknowledged precursors to his approach as far back as Fisher (1935) and Neyman (1923). Extensions to multivalued and continuous treatments have been made by various authors (e.g. Imbens, 2000; Hirano and Imbens, 2004). A key feature of the potential outcomes framework is that it is essentially nonparametric. It does not specify the parametric form of the models used to represent causal inference, but instead derives identification using only probability and expectation. Of course, an important implication of this is that the potential outcomes framework is applicable whatever form of model is adopted.

3.1. Potential outcomes, causal estimands and challenges of causal inference

To begin, we introduce the key features and challenges of our causal inference problem. We want to infer the effect that a treatment D has on a defined outcome Y. The treatment in question could be binary, $\mathcal{D} \in \{0,1\}$, multivalued, $\mathcal{D} \equiv (d_0, d_1, \ldots, d_m)$, or continuous, $\mathcal{D} \subseteq \mathbb{R}$.

For any level of treatment we define an associated *potential outcome*, a random variable measuring response under different treatment regimes.

Definition 3 (*Potential Outcomes*)**.** *For each unit i we define $Y_i(d)$ as the potential outcome for unit i when exposed to treatment $D_i = d$. The full set of potential outcomes for each unit is then $\mathcal{Y}_i = \{Y_i(d) : d \in \mathcal{D}\}$.*

Thus, for a binary treatment there are two potential outcomes for each unit, $Y_i(1)$ and $Y_i(0)$; for a multivalued treatment there are m outcomes, $Y_i(d_0), Y_i(d_1), \ldots, Y_i(d_m)$, and for a continuous treatment there are potentially infinite outcomes denoted in general by $Y_i(d)$ for d in $\mathcal{D} \subseteq \mathbb{R}$.

We will make use of potential outcomes to calculate the causal effect of D on Y. The aim is to compare the outcome actually observed with other potential outcomes had the treatment taken on a different level. There are three *causal estimands* of possible interest.

Definition 4 (*Individual Causal Effect*)**.** *The individual causal effect (ICE),*

$$\tau_i = Y_i(d) - Y_i(d_0), \quad d \neq d_0,$$

is the difference in outcomes for unit i under treatment level $D_i = d$, relative to reference treatment level $D_i = d_0$.

Definition 5 (*Average Potential Outcome (APO)*)**.** *The average potential outcome,*

$$\mu(d) = \mathbb{E}[Y_i(d)],$$

is the expected outcome in the population had all units been treated at level $D_i = d$.

Definition 6 (*Average Treatment Effect (ATE)*)**.** *The average treatment effect,*

$$\tau(d) = \mu(d) - \mu(d_0) = \mathbb{E}[Y_i(d) - Y_i(d_0)],$$

is the difference in expected outcomes for all units in the population under treatment level $D_i = d$ relative to some other reference level of treatment $D_i = d_0$.

Note that, as expectations over the sample, the APO and ATE measure *population* average causal effects. Note also, that often the reference level of treatment is no treatment, e.g. $d_0 = 0$; and in the case of binary treatments the ATE is defined $\tau(1) = \mathbb{E}[Y_i(1) - Y_i(0)]$, being the difference in expected outcomes under intervention (treatment) and nonintervention (control).

Our ability to identify causal estimands using observational data is determined fundamentally by two key challenges that shape the practice of causal inference.

Challenge 1: Missing Data

Of the full set of potential outcomes for unit i, $\mathcal{Y}_i$, we observe only one element, the *actual outcome*

$$Y_i = \sum_{d \in \mathcal{D}} I_d(D_i) Y_i(d),$$

where $I_d(D_i)$ is the indicator function for receipt of treatment dose d, e.g.

$$I_d(D_i) = \begin{cases} 1 & \text{if } D_i = d. \\ 0 & \text{if } D_i \neq d \end{cases}.$$

Outcomes at all other levels, $d \neq D_i$, are unobserved and we refer to these as *counterfactual outcomes*. Thus, for a binary treatment we observe $Y_i = Y_i(1) I_1(D_i) + Y_i(0)(1 - I_1(D_i))$, but we do not observe the joint density $f(Y_i(0), Y_i(1))$, since the two outcomes never occur together. Holland (1986) refers to the challenge of missing data as the 'fundamental identification problem of causal inference'.

An immediate implication of the missing data challenge is that we cannot observe ICEs. For example, with respect to our case studies: if a link has a speed camera at time t we do not observe it without one at t (**CS1**), if household h is resident in zone z at time t we do not observe it resident in other zones at t (**CS2**), if new signalling is not assigned we do not observe the metro line with an upgrade (**CS3**), and if a location falls with an LEZ at t_0 we do not observe it without LEZ status at $t > t_0$ (**CS4**).

Note that in addition to being unobservable, ICEs are also not identifiable statistically because the available data do not provide enough information. For this reason, ICEs tend not to feature as targets of inference.

The key insight of the potential outcomes approach is that if we focus on estimating average causal effects over the population, rather than ICEs, then identification is possible even in the presence of missing data. In fact, if the treatment is *randomly assigned across the population*, then identification of quantities such as APOs and ATEs is straightforward because randomisation implies unconditional independence of treatment assignment and outcome. The argument is as follows.

Theorem 1 (*Identification of the APO Under Unconditional Independence*)**.** *A random treatment assignment implies unconditional independence of the form*

$$Y_i(d) \perp\!\!\!\perp I_d(D_i)$$

for all $d \in \mathcal{D}$. If the potential outcomes are unconditionally independent of the treatment assignment, then APOs are identified since

$$\mu(d) = \mathbb{E}\left[Y_i(d)\right] = \mathbb{E}\left[Y_i(d) | I_d(D_i)\right] = \mathbb{E}\left[(Y_i | D_i = d)\right].$$

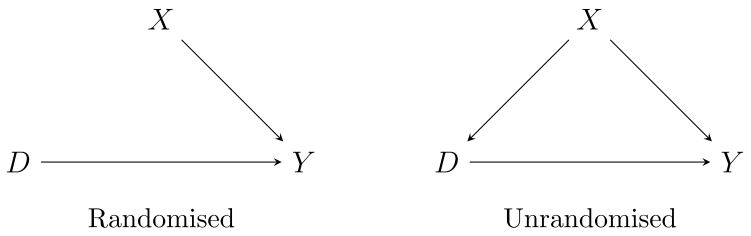

Fig. 1. Directed acyclic graph of randomised and non-randomised treatment assignment.

Unconditional independence allows us to condition potential outcomes on observed treatment statuses because missing data are missing at random. Theorem 1 justifies estimation of APOs and the ATE via sample means using

$$\hat{\tau}(d) = \frac{\sum Y_i \cdot I_d(D_i)}{\sum I_d(D_i)} - \frac{\sum Y_i \cdot I_0(D_i)}{\sum I_0(D_i)}, \tag{1}$$

and consistency of estimation can be established by simply invoking a weak law of large numbers (WLLN): e.g. $n^{-1} \sum_i Y_i I_d(D_i) \xrightarrow{p} \mathbb{E}\left[Y_i(d)\right]$. However, while identification, and consequently consistency, can be established under a random treatment assignment, in reality treatments are often not randomly assigned. If so, theorem 1 no longer applies and identification and consistency fail.

Challenge 2: Non-random assignment and the problem of confounding

This bring us to the second challenge in attaining causal identification in the observational setting: non-random assignment and confounding. These concepts are related. We first define confounding.

Definition 7 (*Confounding*)**.** *A treatment assignment is said to be confounded if characteristics of the data generating process create unconditional dependence between potential outcomes and treatment assignment such that*

$$Y_i(d) \not\perp\!\!\!\perp I_d(D_i)$$

for $d \in D$. Confounding is present when a treatment assignment mechanism induces dependency with covariates X_i, which are themselves important in determining outcome Y_i.

A confounder, therefore, is simply a random variable that influences the outcome of interest, but that is also important in determining assignment to treatment. It is worth noting that confounding can arise dynamically, with past outcomes or treatments of units serving as baseline confounders (e.g. reverse causality).

Fig. 1 shows a simple graphical comparison of randomised (unconfounded) and non-randomised (confounded) treatment assignments.

Under a random assignment unit characteristics X_i have no influence on the treatment received (i.e. on D_i), and consequently there are no *systematic* differences between units receiving different levels of the treatment. As explained above, the property of unconditional independence allows us to treat unobserved potential outcomes much like data that are missing at random and form consistent estimators of APOs and ATEs via sample means.

Under non-randomisation, however, assignment of the treatment depends on covariates X_i, which are themselves important in determining outcome Y_i. As shown in the DAG above, this creates two sources of association between D and Y: (1) via the causal (or direct) path $D \rightarrow Y$ and (2) via the back door (or indirect) path $D \leftarrow X \rightarrow Y$. Thus, some part of the overall association between treatment and outcome could be attributed to X_i rather than D_i. Essentially, the problem that arises with non-randomisation is that the treatment assignment mechanism is determining missingness: that is, which potential outcomes are observed and which are missing. Under these circumstance X_i features as *confounders*, and simple comparisons of mean responses by treatment status will not in general reveal a 'causal' effect because treated and control responses differ irrespective of treatment status.

This is precisely the setting we often find ourselves in when analysing the effects of transport interventions and exposures in the observational setting. For example, with regard to our case studies, the following sources of confounding could be hypothesised.

CS1: Speed cameras (D) may be assigned according to a set of criteria (X) that simultaneously influence collisions rates (Y). These could include: perceived risk (past collision rates), length of road link, speed distribution of link, and history of speed limit violations.

CS2: Self-selection in household location (D) may cause confounding by inducing heterogeneity in household characteristics (X) which also determine car use (Y).

CS3: Signalling upgrades (D) may be assigned to lines that are busier, have older technology, and a history of delay (X). Characteristics X are also relevant for journey times (Y).

CS4: Low emission zones (D) may be assigned to locations with poor air quality, dense land use, and higher shares of public transport (X), which in turn affect air quality (Y).

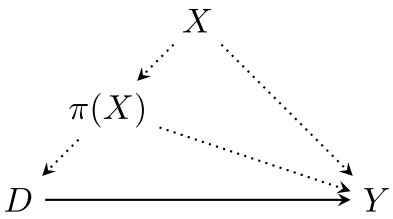

Fig. 2. DAG showing conditional independence of Y and D given X or $\pi(X)$.

Simply ignoring confounding can lead to incorrect inferences on causal effects, and ultimately, yield misleading insights for transport policy and practice.

The key issue for identification of causal effects lies in successful adjustment for confounding. The next two sections of the paper cover the statistical theory underpinning identification and estimation. The focus in these sections is on core concepts and methods.

4. Identification of causal effects via potential outcomes

While the two challenges of causal inference defined above have consequence for identification, consistent estimates of APOs and ATEs can still be obtained under the potential outcomes framework. In this section of the paper, we define the conditions under which causal identification can be achieved in the presence of confounding. There are some important differences between the identification conditions for binary versus multivalued/continuous treatments, so we will describe both.

Within the potential outcomes framework there are three key assumptions required for valid APO and ATE identification in the presence of confounding. These are as follows.

Assumption 1 (*Conditional Independence*)**.** *The potential outcomes for unit i must be conditionally independent of the treatment assignment given a (sufficient) set of observed pre-treatment covariates X_i. For binary treatments the assumption requires that*

$$Y_i(0), Y_i(1) \perp\!\!\!\perp I_1(D_i)|X_i, \tag{2}$$

and for multivalued or continuous treatments Imbens (2000) and Hirano and Imbens (2004) introduce the concept of weak conditional independence which can be stated as

$$Y_i(d) \perp\!\!\!\perp I_d(D_i)|X_i \quad \text{for all} \quad d \in \mathcal{D}. \tag{3}$$

The key difference between the binary and non binary assumptions is that in the latter conditional independence is required to hold for each value of the treatment (i.e. pairwise), but not joint independence of all potential outcomes.

The conditional independence assumption (CIA) thus requires that, conditional on some set of pre-treatment covariates, assignment to treatment does not depend on the outcome. If X_i is sufficient for this to hold then we can in effect mimic, with observational data, the assignment that would occur under randomisation in which the treatment is allocated independently of pre-treatment characteristics.

This principle is illustrated in the DAG shown in Fig. 2. Conditioning on X, or some function of X, $\pi(X)$ say, blocks the indirect path $D \leftarrow X \rightarrow Y$. This nullifies confounding, essentially recapitulating randomisation, and ultimately identifies the direct causal effect $D \rightarrow Y$.

Assumption 2 (*Common Support*)**.** *The support of the conditional distribution of X_i given a particular treatment status should overlap with that of X_i given any other treatment status. For binary treatments this requires that the probability of assignment to the treatment lies strictly between zero and one*

$$0 < \Pr(I_1(D_i) = 1|X_i = x) < 1, \ \forall \ x. \tag{4}$$

For multivalued or continuous treatments we require common support by treatment status in the covariate distributions within some region of dose $C \subseteq \mathcal{D}$. A sufficient condition is that for any subset of C, say $\mathcal{A} \subseteq C$,

$$\Pr(D_i \in \mathcal{A}|X_i = x) > 0, \ \forall \ x \tag{5}$$

The intuition behind the common support, or overlap, assumption is that if some sub-populations observed in X_i have zero probability of receiving (or not receiving) a treatment, then it does not make sense in these cases to talk of a treatment effect since the counterfactual will not exist in observed data.

Assumption 3 (*Stable Unit Treatment Values*)**.** *The relationship between observed and potential outcomes must comply with the Stable Unit Treatment Value Assumption (SUTVA) (e.g. Rubin, 1978, 1980, 1986, 1990), which requires that the observed response under a given treatment allocation is equivalent to the potential response under that treatment allocation. For binary treatments we require that*

$$Y_i = I_1(D_i)Y_i(1) + (1 - I_1(D_i))Y_i(0) \tag{6}$$

for all $i = 1, \dots, N$. For multivalued or continuous treatments we require

$$Y_i \equiv I_d(D_i)Y_i(d) \tag{7}$$

for all $d \in \mathcal{D}$, for all $Y_i(d) \in \mathcal{Y}_i$, and for $i = 1, \dots, N$.

The SUTVA imposes two conditions. First, that the outcome for each unit be independent of the treatment status of other units, or in other words, there should be no interference in treatment effects across units (Cox, 1958). Second, that there are no different versions of the treatment. The SUTVA is generally satisfied when the units are physically distinct and have no means of contact. Violations of the assumption can occur when proximity of units allows for contact and this presents a particular concern for transport applications (e.g. Graham et al., 2013).

The three assumptions defined above, which are together referred to by Rosenbaum and Rubin (1983) as *strong ignorability*, permit identification of APOs and ATEs for non randomly assigned treatments. This is stated formally in the following theorem.

Theorem 2 (*Identification of the APO and ATE under strong ignorability*)**.** *Under strong ignorability APOs and ATEs can be identified by conditioning on X and integrating over the covariate distributions to capture the marginal causal effect.*

$$\begin{aligned}\tau(d) = \mathbb{E}[Y(d) - Y(0)] &= \int \mathbb{E}[Y(d) - Y(0)|X]dF(X) \\ &= \int \mathbb{E}\left[Y|X, I_d(D)\right] - \mathbb{E}\left[Y|X, I_0(D)\right] dF(X)\end{aligned}$$

Note that the ATE is defined as an expectation over covariates X. If we do not take this expectation, but instead simply use the integrand, we obtain an estimate of the causal effect of D within strata of X. In other words, we get the *conditional treatment effect*, that is the ATE for units with characteristics $X = x$. By integrating X out of this distribution we get the average causal intervention distribution. Proofs of theorem 2 in the binary, and continuous/multivalued, treatment settings are provided in Appendix A.1.

Theorem 2 is important because it states clearly how we can uniquely determine causal quantities with observational data, and furthermore, it applies to non-randomly assigned treatments. In the next two sections, we turn to methods that draw on this principles to estimate causal parameters. These fall into one of two classes depending on the data available to represent $X = (X_1, \dots, X_k)$

(1) With X observed: we generate causal evidence by conditioning on X, or some function of X, and assume ignorability under the CIA.
(2) With X completely or partially unobserved: we utilise exogenous variation (e.g. via natural experiments) for identification, and make assumptions other than the CIA.

Following the discussion, we will illustrate each class of method for empirical problems typical of those encountered in transport research.

5. Methods for treatment effect estimation under ignorability

Under strong ignorability, estimation of APOs and ATEs can be achieved via outcome regression (OR) models, Propensity Score (PS) models, and mixed or Doubly Robust (DR) models. To emphasise how these approaches relate, we adopt the notation of Tsiatis and Davidian (2007) and write joint densities of the observed data in the form

$$f_Z(z) = f_{Y|D,X}(y|d, x)f_{D|X}(d|x)f_X(x). \tag{8}$$

OR models focus on $f_{Y|D,X}(y|d, x)$, PS models on $f_{D|X}(d|x)$ and DR models use both. We now explain the three key estimation approach in turn.

5.1. Outcome regression (OR) models

OR approaches are widely used in transport analysis and will likely be familiar to most, so they can be covered here in outline. In the causal inference setting, and with reference to (8), the OR model leaves $f_{D|X}(d|x)$ and $f_X(x)$ unspecified, and explicitly models the conditional expectation function $\mathbb{E}[Y_i|D_i, X_i]$. The focus on this conditional expectation is motivated by theorem 2, which shows that identification can be achieved by averaging over this function.

The conditional expectation, or mean response, could be specified as a linear regression model, a Generalised Linear Model (GLM), a Generalised Linear Mixed Model (GLMM), or some parametric or semiparametric variant thereof. If the OR model, denoted $\Psi^{-1}\{m(D_i, X_i; \xi)\}$, for known link function Ψ, is correctly specified for the conditional expectation, then consistent estimation of ATE can be achieved using

$$\widehat{\tau}_{OR}(1) = \frac{1}{n}\sum_{i=1}^{n}\left[\Psi^{-1}\left\{m(1, X_i; \widehat{\beta})\right\} - \Psi^{-1}\left\{m(0, X_i; \widehat{\beta})\right\}\right],$$

in the binary case and

$$\widehat{\tau}_{OR}(d) = \frac{1}{n}\sum_{i=1}^{n}\left[\Psi^{-1}\left\{m(d, X_i; \widehat{\beta})\right\} - \Psi^{-1}\left\{m(0, X_i; \widehat{\beta})\right\}\right],$$

in the case of continuous treatments.

In other words, simply taking the mean of the predicted values of the OR model, with D held fixed at some value d, gives a consistent estimate of the APO $\mu(d) = \mathbb{E}[Y_i(d)]$.

Note, however, that consistency of OR based adjustment requires that the observed X covariate vector adequately represents *all* sources of confounding, and thus is sufficient to invoke conditional independence. In practice, the true DGP is not known, and there is in fact no way to objectively test whether the CIA has been satisfied or not. Theory provides a basis upon which sources of confounding (and simultaneity) can be hypothesised. If our assumed DGP indicates the existence of confounders that we cannot measure, then we cannot achieve identification via OR adjustment with measured covariates alone.

There are some familiar, and widely used, approaches that can add robustness to the CIA assumption in the OR setting. For example, longitudinal (or panel) data methods attempt to address violation of the CIA by adjusting for unobserved time-invariant unit level characteristics that are believed to induce confounding (see Wooldridge, 2010). Such approaches are to be recommended over cross-sectional specifications when the available data allow them to be used. However, if we believe that unmeasured confounding is present, beyond what can reasonably be captured via panel data adjustments, then we must turn to methods that offer additional robustness. Several such approaches are discussed in Section 6 of the paper.

5.2. Propensity score (PS) models

The focus of OR models is on the relationship between outcome and treatment conditional on covariates. Alternatively, we can model causal relationships by studying the relationship between the treatment assignment and the covariates. Thus, with reference to (8) above, we leave $f_{Y|D,X}(y|d,x)$ and $f_X(x)$ unspecified, but assume a model for $f_{D|X}(d|x)$. This model is used to form *Propensity Scores* (PS), which measure the probability of assignment to treatment given the set of observed pre-treatment covariates (introductions to the main ideas underpinning the PS are given in Joffe and Rosenbaum, 1999; Rosenbaum, 1999; Rubin, 2006).

Definition 8 (*Propensity Score*). *The propensity score measures the conditional probability of assignment to treatment given pre-treatment covariates X_i. For binary treatment the PS is defined*

$$\pi\left(D_i = 1|X_i\right) = \Pr(I_1(D_i) = 1|X_i = x)$$

and for multivalued or continuous treatments

$$\pi\left(d|X_i\right) = f_{D|X}(d|x_i)$$

Before discussing the main estimation methods based on the PS, we first set out the key identification issues.

5.2.1. Identification via propensity scores

An important result, due to Rosenbaum and Rubin (1983), is that the CIA (i.e. Eqs. (2) and (3)) can be restated by replacing the covariate vector X_i with the scalar PS. Rosenbaum and Rubin (1983) proved this result in the case of binary treatments, and Imbens (2000) and Hirano and Imbens (2004) generalise the PS to cover the case of multivalued and continuous treatments. The reason the CIA can be established conditional on the PS, rather than full covariate vector, is because the PS has a *balancing property*, in the sense that it balances the distribution of the observed covariates within strata of the sample that have the same PS. Balancing does not eliminate heterogeneity in X_i, but renders it such that on average units with the same PS can be treated as observationally equivalent (for a formal account see Lemma 1 and Corollary 1 and associated proofs in Appendix A.2).

Thus, if strong ignorability holds, and the model is correctly specified, the PS effectively adjusts for confounding and can thus be used for causal identification. This argument is summarised formally in the theorem below (for a proof, see Appendix A.3).

Theorem 3 (*Identification Given the Propensity Score Under Strong Ignorability*). *Suppose that assignment to the treatment is unconfounded conditional on pre-treatment characteristics X_i. Then we can write*

$$\mathbb{E}[Y_i(d)|I_d(D_i), \pi(d|X_i)],$$

as the conditional mean of the outcome given the treatment level $D = d$ and the PS. If strong ignorability holds then the expectation of $\mathbb{E}[Y_i(d)|I_d(D_i), \pi(d|X_i)]$ over X_i provide an expression for the APO that is identified

$$\mu(d) = \mathbb{E}_{X_i}\left[\mathbb{E}[Y_i(d)|I_d(D_i), \pi(d|X_i)]\right] = \mathbb{E}_X\left[\mathbb{E}\left[Y_i|D_i = d, \pi(d|X_i)\right]\right].$$

5.2.2. Estimation via propensity scores

Given that identification can be achieved via the PS, we can construct PS based estimators for causal quantities. PSs are not observed but are calculated by estimating the relationship between D and X using a regression model

$$\mathbb{E}[D_i|X_i] = \Psi^{-1}\{m(D_i, X_i; \alpha)\}$$

for link function Ψ, regression function $m(\cdot)$, and unknown parameter vector α. The estimated parameters of this model are then used to compute propensity scores, $\pi(D_i|X_i; \widehat{\alpha})$.

The estimated PS, $\pi(D_i|X_i;\widehat{\alpha})$, can then be used to form various nonparametric and semiparametric estimators. A key advantage in using the PS is that it avoids the need to condition on a potentially high dimensional covariate vector, and it is this dimensions reducting property that allows for effective implementation of a number of flexible estimators. Another advantage of the PS is that it is highly effective in isolating the region of common support, a task that is difficult using multiple covariates (for discussion see Joffe and Rosenbaum, 1999).

Below we briefly describe the most commonly used PS estimators which are realised via weighting, regression, and matching. For further details see Imbens and Wooldridge (2009), Imbens and Rubin (2015), Graham et al. (2014).

Inverse propensity score weighting

Inverse PS weighting (IPW) provides a very simple estimator which is based on the principle that the APO, $\mathbb{E}[Y_i(d)]$, can be estimated using

$$\widehat{\mu}_{IPW}(d) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_d(D_i)\cdot Y_i}{\pi(d|X_i;\widehat{\alpha})}\right], \tag{9}$$

which by the WLLN is consistent if the PS model is correctly specified,e.g.

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_d(D_i)\cdot Y_i}{\pi(d|X_i;\widehat{\alpha})}\right] \xrightarrow{p} \mathbb{E}\left[\frac{I_d(D_i)\cdot Y_i}{\pi(d|X_i;\widehat{\alpha})}\right]$$

and by the central limit theorem, $\sqrt{N}$ asymptotically normally distributed (e.g. Rosenbaum, 1987; Joffe and Rosenbaum, 1999; Hirano et al., 2003). A proof for identification of $\mu(d)$ via IPW is provided in Appendix A.4.

From (9) IPW can be use to estimate an ATE for binary treatments of the form

$$\widehat{\tau}_{IPW}(1) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_1(D_i)\cdot Y_i}{\pi(D_i=1|X_i;\widehat{\alpha})} - \frac{(1-I_1(D_i))\cdot Y_i}{1-\pi(D_i=1|X_i;\widehat{\alpha})}\right], \tag{10}$$

and for continuous and multivalued treatments of the form

$$\widehat{\tau}_{IPW}(d) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_d(D_i)\cdot Y_i}{\pi(D_i=d|X_i;\widehat{\alpha})} - \frac{I_0(D_i)\cdot Y_i}{\pi(D_i=0|X_i;\widehat{\alpha})}\right], \tag{11}$$

The IPW estimator originated in Horvitz and Thompson (1952). The principle it invokes revolves around creation of a *pseudo-sample* to simulate random assignment by using the conditional probabilities to mimic the sample representation of units that would occur under randomisation. The intuition is as follows. We know that under a random assignment the probability of assignment to treatment is uniform across unit in the sample. By measuring how baseline characteristics cause deviations from this uniform probability, we can effectively weight units to alter their representation in the sample such that we reproduce an assignment that is, for all intents and purposes, *as good as random*.

This same principle is used in so called Marginal Structural Models (MSMs) (see Robins, 1999b,a; Robins et al., 2000a), which make use of inverse probability weighting in a slightly different form, but are otherwise conceptually very similar. MSMs, along with G-computation (e.g. Robins, 1986) are useful in estimation of dynamic causal relationships, in which time varying covariates act simultaneously as confounders and as intermediate variables.

Note that the consistency of ATE estimation under IPW relies on the PS model being correctly specified. That is, the estimated PSs must provide an accurate measure of the conditional probability of assignment to treatment.

Regression on the propensity score (PSR)

Since the CIA can be established conditional on the PS rather than covariate vector X_i, PSs can be substituted into a regression model in place of X_i. Thus, adapting the OR model discussed above, we can estimate an ATE for treatment dose $D_i = d$ using

$$\widehat{\tau}_{PSR}(d) = \frac{1}{n}\sum_{i=1}^{n}\left[\Psi^{-1}\left\{m(d,\pi(d|X_i;\widehat{\alpha});\widehat{\gamma})\right\} - \Psi^{-1}\left\{m(0,\pi(0|X_i;\widehat{\alpha});\widehat{\gamma})\right\}\right] \tag{12}$$

where γ is a set of parameters to be estimated and $\widehat{\alpha}$ are parameters of the PS model estimated in a prior step. The logic underpinning PSR identification simply follows from the SUTVA and the balancing property of the PS, and is as follows

$$\mu(d) = \mathbb{E}[Y_i(d)] = \mathbb{E}_X\left[\mathbb{E}(Y_i(d)|X_i)\right] = \mathbb{E}_X\left[\mathbb{E}(Y_i(d)|\pi(d|X_i;\alpha))\right] = \mathbb{E}_X\left[\mathbb{E}(Y_i|I_d(D_i),\pi(d|X_i;\alpha))\right].$$

A potential advantage of PSR is that by reducing the dimensionality of the model, that is by using a scalar PS in place of covariate vector X_i, estimation via parametric polynomial models or, semiparametric spline models, could produce a good approximation to the conditional expectation. However, a potential disadvantage, noted by Imbens and Wooldridge (2009), is that since the PS itself is purely a statistical quantity, with no substantive meaning, interpretation of the regression results, and thus evidence on the nature of confounding, could be somewhat opaque.

Matching on the propensity score

A procedure that is similar in principle to blocking is that of matching on the PS. The principle behind matching methods is to find units for the treated and control sub-samples that are as similar to each other as possible. The matching estimator has been used extensively in the literature (see for example Rosenbaum, 1989, 2002; Rubin, 1973a,b, 1979; Rubin and Thomas, 1996, 2000; Abadie and Imbens, 2002).

In its simplest form, matching means fitting an algorithm to identify treated and control units that are similar across all covariates. The PS provides an ideal value upon which to match due to its balancing property. In many applications, a one-on-one mapping of treatment to control units is attempted via the PS, other approaches match a small number of 'neighbours' to approximate the unobserved outcome for a given unit. This gives the simple matching estimator

$$\hat{\tau}_M = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i(1) - \hat{Y}_i(0)), \tag{13}$$

where

$$\hat{Y}_i(1) = \begin{cases} Y_i & \text{if } I_1(D_i) = 1 \\ \frac{1}{M}\sum_{j\in\mathcal{J}_{M(i)}} Y_j & \text{if } I_1(D_i) = 0 \end{cases},$$

and

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } I_1(D_i) = 0 \\ \frac{1}{M}\sum_{j\in\mathcal{J}_{M(i)}} Y_j & \text{if } I_1(D_i) = 1 \end{cases},$$

where $\mathcal{J}_{M(i)}$ is the matched set of $1,\dots,M$ control units for unit i.

The properties of simple matching estimators have been studied extensively by Abadie and Imbens (2002) and Abadie and Imbens (2006). They show that if the dimensions of the continuous covariates is greater than two and if the matching is not exact, then the matching estimator is not consistent with bias due to matching discrepancies of order $O(n^{-1/k})$, where K is the dimension of the continuous covariates. Abadie and Imbens (2006) also show that matching estimators are generally not efficient. However, matching can work well in certain settings, and particularly when combined with other procedures such as regression. A courser version of matching, called PS stratification or blocking, can also be used to partition into strata within which there is little variation in the PS (e.g. Rosenbaum and Rubin, 1983).

5.3. Doubly robust (DR) models

Consistency of causal estimation under PS or OR models requires correct model specification. If the models are misspecified in some way, then valid inference can no longer be guaranteed. It is this issue of susceptibility to misspecification that has given rise to a class of doubly robust (DR) models, which combine OR and PS models to introduced an enhanced robustness property.

With reference to (8) above, under a DR approach, we leave $f(x)$ unspecified and assume a model for both $f_{Y|X}(y|x)$ and $f_{D|X}(d|x)$, and form an estimator that combines both OR and PS models. The key feature of DR estimators is that APO and ATE estimates are consistent and asymptotically normal when either the OR or the PS model are correctly specified, but we do not require both models to be correct (e.g. Robins, 2000; Robins et al., 2000b; Robins and Rotnitzky, 2001; van der Laan and Robins, 2003; Lunceford and Davidian, 2004; Bang and Robins, 2005; Kang and Schafer, 2007).

The motivation for doubly-robust estimation, therefore, is that the analyst effectively has two chances at getting a model specification right. In practice, double robustness is typically achieved by weighting or augmenting the OR model with an inverse PS covariate, often referred to as a 'clever covariate'. A statement of the DR property is as follows.

Theorem 4 (*Double-robustness*). *An APO or ATE estimator, formed by combining an OR model $\Psi^{-1}\{m(X_i, D_i;\beta)\}$ and a PS model $\hat{\pi}(D_i|X_i;\hat{\alpha})$, is doubly-robust if the estimator is consistent when either $\Psi^{-1}\{m(X_i, D_i;\beta)\}$ is correctly specified for $\mathbb{E}[Y_i|D_i, X_i]$, or, $\hat{\pi}(D_i|X_i;\hat{\alpha})$ is correctly specified for $\pi(D_i|X_i)$.*

Thus, for treatment level $d \in \mathcal{D}$ we define $\hat{\mu}_{DR}(d)$ by

$$\hat{\mu}_{DR}(d) = \frac{1}{n}\sum_{i=1}^{n}\left[\Psi^{-1}\left\{m(d, X_i;\hat{\beta})\right\} + \frac{I_d(D_i)}{\hat{\pi}(d|X_i;\hat{\alpha})}\left[Y_i - \Psi^{-1}\left\{m(d, X_i;\hat{\beta})\right\}\right]\right], \tag{14}$$

Which we form by assuming a form for $f_{D|X}(d|x_i,\alpha)$, and estimating the parameter $\hat{\alpha}$ from a regression model using the observed treatment doses D_i and covariates X_i.

We can view the DR APO (14) as a predicted estimating equation with residual bias correction (e.g. Kang and Schafer, 2007). Such an estimating equation can be formed by weighting or augmenting the regression model $\Psi^{-1}\left\{m(X_i, D_i;\beta)\right\}$ with $\widehat{\kappa}_i(d, X_i) = I_d(D_i)/\hat{\pi}(d|X_i;\hat{\alpha})$. Essentially, the argument is as follows. The augmented regression model wil be consistent if the OR model $\Psi^{-1}\left\{m(X_i, D_i;\beta)\right\}$ is correct for $\mathbb{E}(Y(d)|X)$ because the inclusion of the covariate $\widehat{\kappa}_i(D_i, X_i)$ simply adds noise to the predicted values, but leaves the consistency and asymptotic normality of the estimates unchanged. If the OR model is incorrectly specified, but the PS is correctly specified, the model will still be consistent because inclusion of the inverse PS gives rise to estimating equations that effectively correct for the bias in approximating $Y_i(d)$ using $\Psi^{-1}\left\{m(X_i, D_i;\beta)\right\}$ (for details see Bang and Robins, 2005; Tsiatis, 2006; Kang and Schafer, 2007; Graham et al., 2016). For a proof of the DR property in the case of multivalued treatments see Appendix A.5.

The bias correcting estimating equations of the DR model can be derived via a number of standard regular asymptotically linear estimators. For instance, Maximum Likelihood Estimation (MLE), Maximum Quasi-Likelihood (MQL), Restricted MLE (REML) for linear mixed models (LMMs), and Penalised Quasi-Likelihood (PQL) for generalised linear mixed models (GLMMs) all provide estimating equations of the form

$$\sum_{i=1}^{n}\widehat{\kappa}_i(D_i, X_i)\frac{1}{\phi}\frac{\partial\left[\Psi^{-1}\left\{m\left(D_i, X_i, \widehat{\kappa}_i(D_i, X_i);\xi\right)\right\}\right]}{\partial\xi^{\mathsf{T}}}\left[Y_i - \Psi^{-1}\left\{m\left(D_i, X_i, \widehat{\kappa}_i(D_i, X_i);\xi\right)\right\}\right] = 0, \tag{15}$$

where $\phi_i \equiv \phi(D_i, X_i)$ is a working conditional variance for Y_i given (D_i, X_i). Following the approach introduced by Scharfstein et al. (1999), a DR estimate of $\tau(d)$ can then be obtained as

$$\hat{\tau}_{DR}(d) = \frac{1}{n}\sum_{i=1}^{n}\left[\Psi^{-1}\left\{m\left(d, X_i, \hat{\kappa}_i(D_i, X_i); \hat{\xi}\right)\right\} - \Psi^{-1}\left\{m\left(0, X_i, \hat{\kappa}_i(D_i, X_i); \hat{\xi}\right)\right\}\right].$$

In this way, DR models add an additional layer of robustness in the sense that we only have to get one of the two component model right. Of course, it is always possible to get both the OR and PS models wrong, and moreover, if the measure of X available in the observed data is insufficient to guarantee the CIA, then both the OR and PS models will fail and DR estimation will not improve matters. For a Bayesian take on the DR approach see Graham et al. (2016).

5.4. Variance estimation

The estimators described above vary in nature and form. Some are based on particular parametric forms while others are based on non-parametric identification and mild functional form assumptions. In addition, some estimators involve multiple stages of calculation often with sequential plug-in steps. Under correct model specification, the APO/ATE estimates derived from these approaches are typically consistent and asymptotically normal (CAN) in the usual sense, e.g. for parameter τ

$$\begin{aligned}\left(\hat{\tau}_n - \tau\right) &\xrightarrow{p} 0\\ \sqrt{n}\left(\hat{\tau}_n - \tau\right) &\xrightarrow{d} \mathcal{N}\left(0, \sigma_\tau^2\right),\end{aligned}$$

with zero asymptotic bias. It follows that standard errors and standard confidence intervals can be used for testing hypotheses; and that in general, variance estimation is justified by asymptotic theory, either using large sample approximations to the asymptotic variance via the delta method, or via bootstrap approximation.

6. Methods for estimation given a non-ignorable treatment assignment

The validity of the estimation methods discussed in the previous section requires us to maintain that strong ignorability holds. In practice, this is often untenable, either because there are insufficient measured covariates to justify the CIA, or because other sources of endogeneity are at play, such as reverse causality or measurement error, inhibiting a causal interpretation of the data.

There are a number of popular estimators that are used under these conditions to obtain causal estimates of the APO and ATE. Some use additional variables (instruments) to recover exogenous variation in treatments, while others exploit quasi-experimental conditions for identification. Here we cover four of the most commonly used approaches: instrumental variables (IV), difference-in-differences (DID), synthetic control (SC) and regression discontinuity design (RDD).

6.1. Instrumental variables

Instrumental Variables (IV) estimation effectively bypasses the CIA, and directly removes confounding bias, by introducing other observable variable(s) known as *instruments*, which we will denote W.

Consider a general model with additive error, $Y = f(D, X) + e = \mathbb{E}[Y|D, X] + e,\ \mathbb{E}(e) = 0$. We want to know the causal effect of D on Y. Covariates X are exogenous, e.g. $\mathbb{E}(X^{\mathsf{T}}e) = 0$, but insufficient to satisfy the CIA; and D is endogenous, e.g. $\mathbb{E}(D^{\mathsf{T}}e) \neq 0$. In the simple case of a single endogenous treatment variable, we can introduce an instrument W to solve the endogeneity problem. For an instrument to be *valid*, that is, capable of achieving identification in an endogenous model, it must satisfy two conditions.

1. **Exclusion restriction** - the instrument W must have no effect on Y, other than that mediated via D, and it must be uncorrelated with the error in the non-identified model, e.g. $\mathbb{E}\left(W^{\mathsf{T}}e\right) = 0$. In other words, W must be exogenous, and would not feature in the non-identified model.
2. **Relevance** - the instrument must be correlated with the endogenous explanatory variable, conditionally on any other covariates, e.g. $\mathbb{E}\left(W^{\mathsf{T}}D\right) \neq 0$.

The basic principle underpinning IV estimation is that we can achieve identification by using the instruments to enforce orthogonality between the error term and an instrument transformed design matrix. The relationships assumed in IV estimation are shown in Fig. 3.

The defining characteristics of the IV model are as follows: changes in W are associated with changes in D, but do not lead to changes in Y other than through D. W is causally associated with D but *definitely* not with Y, and thus W would not feature in a regression model for Y.

A good way of conceptualising the IV identification principle is as follows. Condition the general model, $Y = \mathbb{E}[Y|D, X] + e$, on W and X and take expectations

$$\begin{aligned}\mathbb{E}[Y|W, X] &= \mathbb{E}\left[\mathbb{E}[Y|D, X]|W, X\right] + \mathbb{E}(e|W, X)\\ &= \int \mathbb{E}[Y|D, X]dF(D|W, X) + \mathbb{E}(e|W, X).\end{aligned}$$

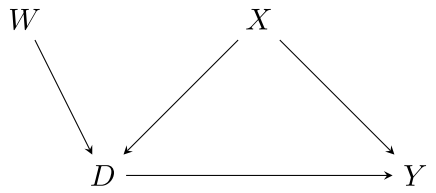

Fig. 3. Relationships in instrumental variables estimation.

Since the exclusion restriction, along with exogeneity of X, implies $\mathbb{E}(e|W,X)=0$, then using an estimate $\hat{F}(D|W,X)$ we could recover consistent a estimate $\hat{\beta}$ from $\mathbb{E}[Y|D,X]$ using observable data. As we will see below, use of IV within the linear model, via two stage least squares, provides an exact solution to the equation above.

The IV estimator is well known and widely used and for that reason we do not provide an extensive description here. We also focus on application of IV within the context of the linear model for a multivalued or continuous treatment variable ($\mathcal{D} \equiv (d_0, d_1, \ldots, d_m)$ or $\mathcal{D} \subseteq \mathbb{R}$), e.g.

$$Y_i = \mathbb{E}[Y_i|D_i,X_i] + u_i = \beta_D D_i + \beta_X X_i + u_i \tag{16}$$

since this is by far the most common setting for IV and exact results can be derived. An up–to-date discussions of IV for binary treatments, non-normal response distributions, and non-linear specifications is provided by Imbens (2014).

Stacking observations for each individual we write the model for the DGP

$$\begin{aligned} Y|D,X &\sim \mathcal{N}(\beta_D D + \beta_X X, \sigma_Y^2) \\ D|X &\sim \mathcal{N}(\alpha X, \sigma_X^2) \end{aligned} \tag{17}$$

We know from the discussion of OR based estimators above that if we have sufficient covariates X to satisfy the CIA, then the OR model

$$Y = \beta_D D + \beta_X X + u. \tag{18}$$

has a causal interpretation. When the CIA fails, due to the presence of unmeasured confounders (e.g. we will assume X is not observed), then this is no longer the case. For instance, say we estimate the model

$$Y = D\beta_D + e \tag{19}$$

instead of the true model (18), then omission of the unobserved confounders, X, will causes problems of omitted variable bias (OVB) and inconsistency in estimation of the ATE. Such a failure amounts to violation of the Gauss–Markov condition that the error term is distributed independently of the regressors: $\mathbb{E}\left(D^{\mathsf{T}}e\right) \neq 0$. Sometime referred to as the population orthogonality condition, failure arises from the fact that $e = \beta_X X + u$, and by the definition of confounding, $\mathbb{E}\left(D^{\mathsf{T}}X\right) \neq 0$.

To see how IV achieve identification, we premultiply our linear model $Y = D\beta_D + u$ by W^{T} to get

$$W^{\mathsf{T}}Y = W^{\mathsf{T}}D\beta_D + W^{\mathsf{T}}e.$$

Taking expectations and using the exclusion restriction $\mathbb{E}[W^{\mathsf{T}}e]=0$, then

$$\mathbb{E}[W^{\mathsf{T}}Y] = \mathbb{E}[W^{\mathsf{T}}D]\beta_D,$$

and we can solve for β_D as

$$\beta_{D_{IV}} = \mathbb{E}[(W^{\mathsf{T}}D)]^{-1}\mathbb{E}[(W^{\mathsf{T}}Y)].$$

Thus, β_D is identified using the IVs since the expectations $\mathbb{E}[(W^{\mathsf{T}}D)]$ and $\mathbb{E}[(W^{\mathsf{T}}Y)]$ can be consistently estimated using observed data (for detail see Appendix A.6).

We have worked with a single endogenous treatment variable and a single instrument, but the IV set up generalises immediately to one in which $\dim W = \dim D \geq 1 = M$. More generally, the so called order condition for IV requires that the number of instruments be greater than or equal to the number of endogenous covariates: e.g. $\dim W = L \geq \dim D = M$. In fact, exogenous covariates can act as instruments in their own right since they satisfy the two IV validity conditions. So we simply need to find as many (or more) valid instruments are there are endogenous covariates. When $L = M$ the model is said to be just-identified. When $L > M$ the model is said to be overidentified.

The IV estimator described above requires that $L = M$. In the overidentified setting we could simply discard some instruments, but this can result in loss of efficiency. Instead, both just identified and over-identified model are usually estimated via two-stage Least Squares (2SLS), which proceeds as follows

1. Regress each column of D on the instrument matrix W and save the predicted values $\hat{D} = W(W^{\mathsf{T}}W)^{-1}W^{\mathsf{T}}D = P_W D$.
2. Regress Y on the predicted values from the first stage: $Y = \hat{D}\beta_D + e$.

The resulting 2SLS estimator is

$$\beta_{D_{2SLS}} = [(D^{\top}W(W^{\top}W)^{-1}W^{\top}D)]^{-1}[(D^{\top}W(W^{\top}W)^{-1}W^{\top}Y)]$$
$$\beta_D + [(D^{\top}W(W^{\top}W)^{-1}W^{\top}D)]^{-1}[(D^{\top}W(W^{\top}W)^{-1}W^{\top}e)],$$

where the second term disappears in expectation due to $\mathbb{E}[W^{\top}e] = 0$. Asymptotic results can be derived by applying a LLN to the sample moments.

The IV estimator can estimate average treatment effects, but where the *average* depends on the choice of instruments, that is, it is conditional on W. We refer to this as the *Local Average Treatment Effect (LATE)*, and it is in general not the same as the *population ATE*.

IV can be used to establish local causal effects under a non-ignorable treatment assignment and is particularly useful when endogeneity via bi-directionality is present. However, it is crucial that the two key assumptions of exogeneity and relevance are met, and in practice valid instruments are hard to find. When instruments are only weakly correlated with the endogenous regressors, or when the instruments themselves are correlated with the error term, IV estimation can produce severely biased and inconsistent estimates. This problem is further confounded by the fact that the available diagnostic statistics do not provide a full proof means for detecting an inadequate instrument specification. To quote Hahn and Hausman (2003), even using standard tests for instrument validity "the researcher may estimate 'bad results' and not be aware of the outcome" (p 118).

6.2. Difference-in-differences

Differences-in-differences (DID) is a 'before and after' treatment effect estimation approach that is applicable when the effect of treatment on units can be represented as a binary variable and observed for treated and control units. It can reveal impacts associated with exposure to an intervention relative to non-exposure (control), but it cannot tell us about impacts by scale or 'dose' of intervention. DID is therefore applicable when $\mathcal{D} \in \{0, 1\}$.

An identification problem will arise for estimation of an ATE in the before-after setting if unobserved confounding is present. This could simply take the form of differences between the treated and untreated units, which affect outcomes, but are also influential in treatment assignment. If present, ignorability will fail, and by extension, identification via OR, PS or DR models is ruled out. In addition to unobserved confounding, there could also be temporal trends that affect the outcome variable due to events unrelated to the treatment.

The DID estimator addresses such potential sources of bias by using information for both treated and control groups in both pre and post treatment periods. In the basic DID model, we model the outcomes, Y_{it}, for units i, $i = (1, 2, \ldots, N)$ in binary time periods $t \in \{0, 1\}$ (with $t = 0$ representing the pre-treatment period and $t = 1$ the post-treatment period) using

$$Y_{it} = \mu + \alpha I_1(D_i) + \delta_t \cdot t + \tau_D \cdot I_1(D_i) \cdot t + \varepsilon_{it}, \tag{20}$$

where υ_{it} is a potentially autoregressive error with mean zero in each time period. The effect of the treatment is captured by the parameter τ_D, which provides the sample counterpart to

$$\tau_D = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \left\{\mathbb{E}[Y_{i,1}|I_1(D_i)] - \mathbb{E}[Y_{i,0}|I_1(D_i)]\right\} - \left\{\mathbb{E}[Y_{i,1}|I_0(D_i)] - \mathbb{E}[Y_{i,0}|I_0(D_i)]\right\}, \tag{21}$$

with least squares estimate

$$\hat{\tau}_D = \left(\overline{Y}_{11} - \overline{Y}_{10}\right) - \left(\overline{Y}_{01} - \overline{Y}_{00}\right),$$

where $\overline{Y}_{11}$ is the sample average outcome for treated units in year 1.

The 'double-differencing' of the DID estimator removes two potential sources of bias. First, it eliminates biases in second period comparisons between the treated and controlled groups that could arise from time invariant confounding characteristics. Second, it corrects for time varying biases in comparisons over time for the treated group that could be attributable to time trends unrelated to the treatment.

There are two key identifying assumptions required of the basic DID model. First, we assume that the treatment assignment and the error ε_{it} are independent

$$\Pr(I_1(D_i)|\varepsilon_{it}) = \Pr(I_1(D_i)),$$

for $t = 0, 1$.

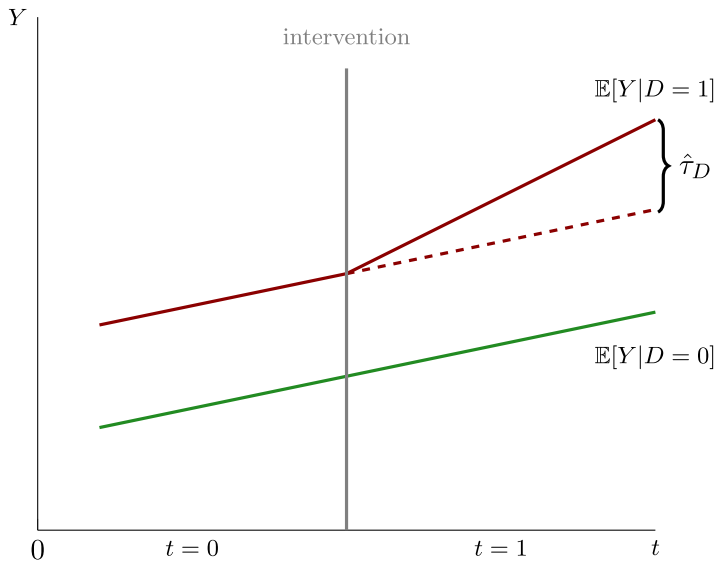

Fig. 4. Difference-in-differences estimation.

Second, we assume that the average outcomes for the treated and control groups would have followed parallel paths over time in the absence of the treatment. If $Y_{it}(0)$ is the outcome that unit i experiences in time t in the absence of treatment, then for binary period $t \in 0, 1$ we make the following assumption.

Assumption 4 (*Unconditional parallel outcomes*)**.** *For identification of treatment effects in the basic DID model it is necessary that the average outcomes for the treated and control groups would have followed parallel paths over time in the absence of the treatment*

$$\mathbb{E}_i \left[Y_{i,1}(0) - Y_{i,0}(0) | I_1(D_i) \right] = \mathbb{E}[Y_{i,1}(0) - Y_{i,0}(0) | I_0(D_i)] \, .$$

The DID estimator is illustrated graphically in Fig. 4.

There are a number of useful extensions of the DID model that can help render the assumptions required for identification more plausible.

1. **DID with covariates** - Covariates (X_i) can be added to the DID model to adjust for heterogeneity in outcome dynamics between treated and control groups. With the vector of covariates included the DID model is

$$Y_{it} = \mu + X_i'\beta + \alpha I_1(D_i) + \delta_t \cdot t + \tau_D \cdot I_1(D_i) \cdot t + \varepsilon_{it}. \tag{22}$$

   The impact of the treatment across the population can be generalised by allowing for interactions between X_i and $I_1(D_i)$. The addition of covariates that are thought to be associated with the dynamics of the outcome variables can be particularly useful in satisfying the identifying restrictions of the DID model. Thus, regarding independence of the error, we can condition on covariates predetermined at $t = 0$ such that

$$\Pr(I_1(D_i)|X_i, \varepsilon_{it}) = \Pr(I_1(D_i)),$$

   is required rather than unconditional independence. Similarly, in the covariate model assumption 4 becomes

$$\mathbb{E}_i[Y_{i,1}(0) - Y_{i,0}(0)|X_i, I_1(D_i)] = \mathbb{E}[Y_{i,1}(0) - Y_{i,0}(0)|X_i, I_0(D_i)],$$

   which may be considerably more plausible.
2. **DID with multiple groups and time periods** - When the number of time periods exceeds 2 the DID model can be reformulated by including a vector of time dummies and an indicator variable which takes a value of 1 for groups and time periods that were subject to the treatment

$$Y_{it} = \mu + X_{it}'\beta + \alpha I_1(D_i) + \sum_{t=1}^{T} \delta_t \cdot I_{T_i}(t) + \tau_D \cdot I_i + \varepsilon_{it}. \tag{23}$$

   where $I_{T_i}(t)$ is an indicator variable for year t and I_i is the group-time treatment indicator.
   For multiple time periods and groups the model is

$$Y_{it} = \mu + X_{it}'\beta + \sum_{g=1}^{G} \alpha_g I_{G_i}(g) + \sum_{t=1}^{T} \delta_t \cdot I_{T_i}(t) + \tau_D \cdot I_i + \varepsilon_{it}, \tag{24}$$

where $I_{G_i}(g)$ is an indicator variable for year membership of group g.

3. **Longitudinal DID** - So far DID has been discussed in a repeated measures setting in which we have a random sample of observations in each time period. If instead the data are for the same units over time the other formulations are possible. For instance, the basic DID model can then be estimated in differences as

$$Y_{i,1} - Y_{i,0} = \delta + \tau_D I_1(D_i) + (\varepsilon_{i,1} - \varepsilon_{i,0}).$$

Furthermore. If we can assume unconfoundedness given lagged outcomes, then a differenced estimator could be $Y_{i,1} - Y_{i,0} = \delta + \tau_{DU} I_1(D_i) + Y_{i,0} + \varepsilon_{it}$. To estimate this model we must either be able to assume that $\mathbb{E}[\varepsilon_{it}|Y_{i,0}] = 0$ or be able to correct for any correlation through, for instance, the use of lagged differences in the outcome variable as instruments in a Generalised Method of Moments estimator.

It is important to note two potential limitations with the DID approach. First, it relies on the strong identifying assumption that the average outcomes for the treated and control groups would have followed parallel paths over time in the absence of the treatment. Second, the model is sensitive to error specification, and in particular, it has been shown that the existence of correlation within groups or over time periods can adversely affect estimation performance (e.g. Bertrand et al., 2004).

6.3. Synthetic control

The synthetic control (SC) approach can be thought of a development of DID which aims to improve comparability of the treatment and control groups (e.g. Abadie and Gardeazabal, 2003; Abadie et al., 2010; Abadie, 2021). This is done by choosing weights for multiple control groups to form a synthetic control unit that is optimally similar to the treated group. The weights can be chosen by considering the distances in averages between treated and untreated groups or by using information on group level covariates.

Let J be the number of available control groups and let K be the number of characteristics we observe for each group. Our objective is to assign an appropriate weight to each control, such that the weighted SC will better approximate the observed characteristics of the treated group. The optimal $J \times 1$ vector of weights $\lambda = (w_1, \dots, w_J)$ is obtained from the following minimisation problem

$$\min_{\lambda}(X_1 - X_0\lambda)^{\mathsf{T}} V (X_1 - X_0\lambda) \tag{25}$$

such that $\sum_{i=1}^{J} w_i = 1$, and where X_1 is a $K \times 1$ matrix containing all the characteristics of the treated group and X_0 is the $K \times J$ vector containing the characteristics of the control groups. As long as X_1 is within the convex hull of X_0 for the donor pool, a weighted average of control units can be used to reproduce $Y_1(0)$ after treatment, $Y_0\lambda$.

The solution to this problem is a vector of optimal weights, $\lambda^*(V)$, which add up to one. The optimal vector depends on V, a $K \times K$ diagonal matrix that assigns weights to each characteristic. We choose the matrix V so that the outcome of interest in the treated group is best replicated by its synthetic counterpart in the period prior to treatment. Let T_p be the number of years observed prior to the treatment, then the problem of choosing V can be specified as follows

$$\min_{V}(Z_1 - Z_0\lambda^*(V))^{\mathsf{T}}(Z_1 - Z_0\lambda^*(V)) \tag{26}$$

where Z_1 is a $T_p \times 1$ vector with a time series of the variable of interest in the treated group, and Z_0 is a $T_p \times J$ vector with a time series of the variable of interest in the control groups.

Once we have derived the SC group, we compare its evolution with respect to the treated group during the years following the treatment. Let T be the number of years observed after the treatment, then we are interested in the gap between the treated group and its synthetic analogue

$$Y_1 - Y_0\lambda \tag{27}$$

where Y_1 is a $T \times 1$ vector with a time series of the variable of interest in the treated group for the post treatment years, and Y_0 is a $T \times J$ vector with a time series of the variable of interest in the control groups for the post treatment years. The SC estimator is illustrated graphically in Fig. 5.

In this way, the SC approach allow us to identify not only a single ATE estimate, but also reveals how the ATE effect evolves over time post-treatment. Clearly, the accuracy of the resulting estimator depends crucially on the choice of weights, and thus on the set of characteristics used to form the weights in the first place. Abadie (2021) and Athey and Imbens (2017) provide a detailed discussion of these issues and how improvements can potentially be made in particular settings.

6.4. Regression discontinuity designs

RDD methods can estimate ATEs under non-ignorability when a given covariate, referred to as the forcing, or running variable, partly or completely determines assignment to the treatment (for reviews of RD see Imbens and Lemieux, 2008; Lee and Lemieux, 2010). Under a so-called 'sharp' RDD design, the conditional probability of receiving the treatment is of size one at some given threshold of the forcing variable, while under a 'fuzzy' design the probability of change at the threshold is less than one. The RDD method exploits this discontinuity in treatment assignment to study the conditional distribution of the outcome either side of the threshold of the forcing variable. A discontinuity in outcome is interpreted as evidence of a causal effect of the treatment.

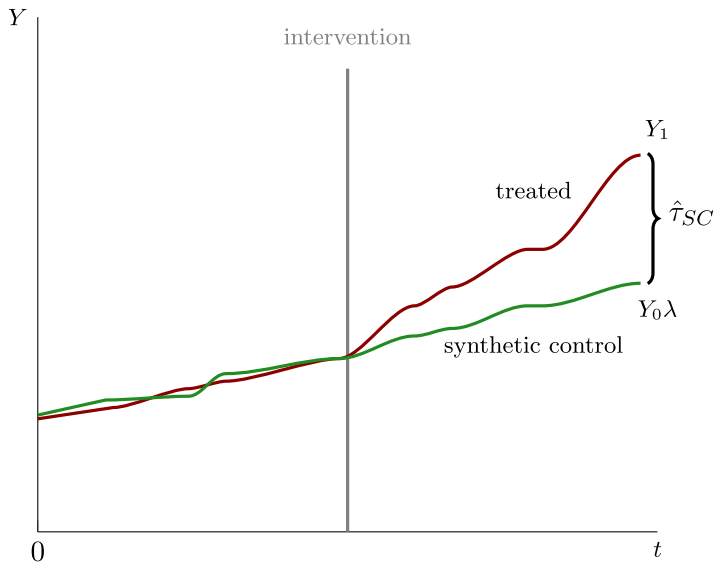

Fig. 5. Synthetic control estimation.

For illustration, we will consider here identification under a sharp RDD design in which treatment assignment is a deterministic function of the forcing variable, which we denote by T. The treatment status of unit i is given by

$$D_i = 1[T_i \geq c]$$

where $1[T_i \geq c]$ is an indicator function that takes a value of one if the statement in brackets is true or zero otherwise. We seek to estimate

$$\tau_{SRD} = \mathbb{E}\left[Y_i(1) - Y_i(0)|T_i = c\right] = \mathbb{E}\left[Y_i(1)|T_i = c\right] - \mathbb{E}\left[Y_i(0)|T_i = c\right],$$

which is equivalent to the population ATE if the treatment effect is constant. We cannot observe both expectations because we have unobserved potential outcomes. Instead, we assume continuity of the expectations in T such that

$$\mathbb{E}\left[Y_i(0)|T_i = c\right] = \lim_{t \uparrow c} \mathbb{E}\left[Y_i(0)|T_i = t\right] = \lim_{t \uparrow c} \mathbb{E}\left[Y_i|T_i = t\right],$$

implying that

$$\tau_{SRD} = \lim_{t \downarrow c} \mathbb{E}\left[Y_i|T_i = t\right] - \lim_{t \uparrow c} \mathbb{E}\left[Y_i|T_i = t\right].$$

The SRD estimator is illustrated graphically in Fig. 6.

The ATE we estimate via the sharp RD design is the difference in the conditional expectation of the outcome either side of the discontinuity. Note that the key identifying assumptions are sharp discontinuity in D and continuity in $Y(0)$ at threshold.

A valid RDD design will provide consistent causal estimates of the ATE without the need to condition on baseline covariates, but it can however be useful to include them anyway to reduce the sampling variability of the estimator and improve precision. The running variable can be any observed covariate, and actually in some applications is simply time.

7. Simulation of causal estimators in transport research

In this section we demonstrate application of causal methods via Monte Carlo simulation of the four case studies described above. We use standard distributions to generate the DGPs, including Normal ($\mathcal{N}$), Bernoulli ($\mathcal{B}$) and Uniform Continuous ($\mathcal{U}$). The units of the generated data are not constrained to impose realism, since the simulations are mainly illustrative. The objective is simply to represent the essence of empirical problems in transport typical of those encountered in practice, and to compare and benchmark the performance of causal estimators.

Each simulation is based on 1000 runs on generated datasets of size 1000. We report mean values (*Av. Est.*) and empirical variances (*Emp. Var.*) of the estimates obtained, and the mean squared error (*MSE*). R code for the simulations is available on request for readers to run and adapt.

CS1: Road safety - simulation of OR, PS & DR models

Our objective is to quantify the causal effect that speed camera presence, $D \in \{0, 1\}$, has on road traffic collisions per lane mile, $Y \subseteq \mathbb{R}$ (e.g. Graham et al., 2019; Li and Graham, 2016). We have data at the link level on D and Y for treated ($D = 1$) and control

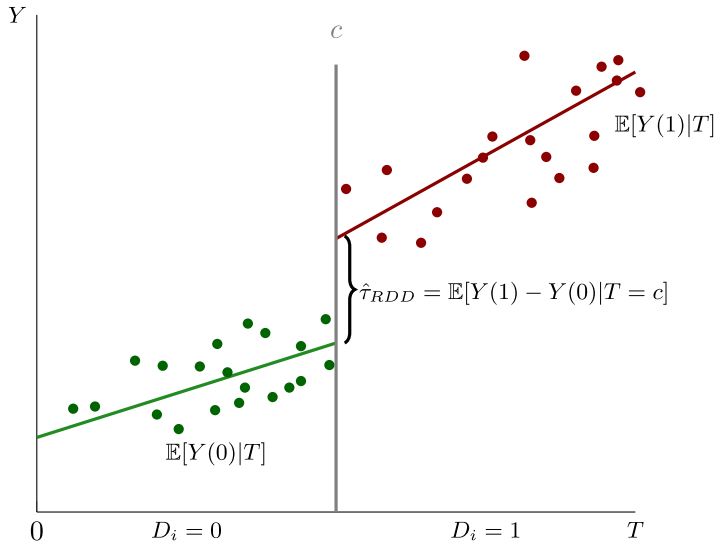

Fig. 6. Sharp regression discontinuity design estimator.

Table 1
Simulation of ATE estimators for CS1 ($\tau = -5.0$).

	Estimator	*Av. Est.*	*Emp. Var.*	*MSE*
1.	$\hat{\tau}_{SM}$	−3.144	0.041	3.488
2.	$\hat{\tau}_{OR1}$	−4.999	0.084	0.084
3.	$\hat{\tau}_{OR2}$	−3.149	0.087	3.512
4.	$\hat{\tau}_{PS1}$	−4.949	2.054	2.054
5.	$\hat{\tau}_{PS2}$	3.697	2.859	78.499
6.	$\hat{\tau}_{DR1}$	−4.944	0.398	0.400
7.	$\hat{\tau}_{DR2}$	−4.977	0.309	0.309
8.	$\hat{\tau}_{DR3}$	−3.125	0.313	3.827

($D = 0$) links. As explained in Section 3, we believe that the assignment of speed cameras is confounded, since they are allocated according to a set of criteria, $X \subseteq \mathbb{R}$, that are also relevant for Y.

We represent the DGP as follows.

$$X \sim \mathcal{N}(0, 10)$$
$$D|X \sim \mathcal{B}(\text{expit}(\alpha_0 + \alpha_1 X))$$
$$Y|D, X \sim \mathcal{N}(\beta_0 + \tau D + \beta_1 X, 5)$$

where we set $\alpha_0 = 2$, $\alpha_1 = 0.5$, $\beta_0 = 10$, $\tau = -5$, $\beta_1 = 0.5$. The true causal effect is the ATE, $\mathbb{E}\left[Y_i(1) - Y_i(0)\right]$, given by parameter $\tau = -5.0$.

The following estimators are simulated

1. $\hat{\tau}_{SM}$ - Naïve model: difference in sample means between treated and control groups.
2. $\hat{\tau}_{OR1}$ - correctly specified OR model: $\mathbb{E}[Y|D, X] = \beta_0 + \tau D + \beta_1 X$.
3. $\hat{\tau}_{OR2}$ - incorrectly specified OR model with covariate X excluded: $\mathbb{E}[Y|D] = \beta_0 + \tau D$.
4. $\hat{\tau}_{PS1}$ - correctly specified IPW estimator using PS model $\mathbb{E}[D|X] = \alpha_0 + \alpha_1 X$.
5. $\hat{\tau}_{PS2}$ - incorrectly specified IPW estimator using random PS values $\hat{\pi}(D|X) \sim \mathcal{N}(\bar{\pi}, 0.5)$, with mean equal to the mean of the true PS and truncated between 0.01 and 0.99.
6. $\hat{\tau}_{DR1}$ - DR model based on an incorrectly specified OR model (e.g. 3.), but with PS weighting based on the correct PS model (e.g. 4.).
7. $\hat{\tau}_{DR2}$ - DR model based on a correctly specified OR model (e.g. 2.), but with weighting based on an incorrect PS model (e.g. 5.).
8. $\hat{\tau}_{DR3}$ - DR model based on an incorrectly specified OR model weighted with weights based on an incorrect PS model (e.g. 3. & 5.).

Table 1 shows the simulation results.

The Naïve (non-causal) ATE estimator ($\hat{\tau}_{SM}$), which ignores confounding, is noticeably biased with a mean value of −3.1 compared to the true τ of −5.0. The correctly specified OR model, $\hat{\tau}_{OR1}$, provides a good estimator for the true value of τ as we

Table 2
Simulation of IV estimator for CS2 ($\beta_D = -1.0$)

	Estimator	*Av. Est.*	*Emp. Var.*	*MSE*
1.	$\widehat{\beta}_{OR_n}$	−0.800	0.001	0.041
2.	$\widehat{\beta}_{OR_t}$	−1.000	0.001	0.001
3.	$\widehat{\beta}_{IV_v}$	−1.000	0.001	0.001
4.	$\widehat{\beta}_{IV_i}$	−0.501	0.002	0.252

would expect from the theory of least squares; but the incorrectly specified OR model fails to satisfy the CIA, and consequently, $\widehat{\tau}_{OR2}$ provides a poor estimator, giving a similar ATE value to the naïve model. A good estimator of τ is achieved via the correctly specified PS model ($\widehat{\tau}_{PS1}$), but when the PS is misspecified (e.g. $\widehat{\tau}_{PS2}$) the average estimate of the ATE is far away from the true value. This tendency of the inverse PS model to fail quite considerably under severe misspecification is well known in the literature (e.g. Kang and Schafer, 2007). Weighting the incorrectly specified OR model with weights $\widehat{\kappa}(D, X)$, based on a correctly specified PS model, (e.g. $\widehat{\tau}_{DR1}$), provides correction for misspecification bias with an average point estimate very close to the true value. The DR2 model also produces valid point estimates because weighting by weights based on an incorrectly specified PS model does not induce bias when the OR model is correct, but it does increase variance. Finally, if both the OR and PS models are wrongly specified, (e.g. $\widehat{\tau}_{DR3}$), the model fails to produce a consistent estimate of the mean ATE.

CS2: Travel behaviour - simulation of IV

Our objective is to quantify the causal effect that urban density, $D \subseteq \mathbb{R}$, has on automobile use, $Y \subseteq \mathbb{R}$ (e.g. Handy et al., 2005). There are a set of unobserved confounders, $X \subseteq \mathbb{R}$, that simultaneously determine D and Y (see Section 3) and their exclusion from the model will induce OVB. To address this we will introduce an instrument, W, which should be correlated with D but not Y. The DGP for this case study is given by

$$\begin{aligned} X &\sim \mathcal{N}(15, 1) \\ W &\sim \mathcal{N}(0, 1) \\ D|X, W &\sim \mathcal{N}(\alpha_0 + \alpha_1 X + \alpha_1\ W, \sigma_D^2) \\ Y|D, X &\sim \mathcal{N}(\beta_0 + \beta_D D + \beta_X X, \sigma_Y^2) \end{aligned}$$

where we set $\alpha_0 = 1$, $\alpha_1 = 0.5$, $\alpha_2 = 1$, $\beta_0 = 1$, $\beta_D = -1$, $\beta_X = 0.5$. The true LATE is given by parameter $\beta_D = -1.0$.

The following estimators are simulated.

1. $\widehat{\beta}_{OR_n}$ - incorrectly specified OR model with covariate X excluded: $\mathbb{E}[Y|D] = \beta_0 + \beta_D D$.
2. $\widehat{\beta}_{OR_t}$ - correctly specified OR model: $\mathbb{E}[Y|D, X] = \beta_0 + \beta_D D + \beta_X X$.
3. $\widehat{\beta}_{IV_v}$ - IV using valid instrument W for $\widehat{D}$ with second stage: $\mathbb{E}[Y|D, X] = \beta_0 + \beta_D \widehat{D}$.
4. $\widehat{\beta}_{IV_i}$ - IV using invalid instrument that is correlated with $u = Y - \mathbb{E}[Y|D, X]$.

The simulation results are shown in Table 2.

The naïve, non-causal OR model, which excludes X and makes no further adjustment, produces a poor estimate of the LATE. The correct OR model, which includes X in the regression, gives an unbiased estimate of the LATE. The IV1 model, estimated via 2SLS using a valid instrument, nullifies the effect of confounding bias and consequently produces a consistent LATE estimate. Note that use of an invalid endogenous instrument, as in IV2, produces a LATE estimate with larger bias than the naïve OR model.

CS3: Network performance - simulation of DID

Our objective is to quantify the causal effect that the presence of a metro signalling upgrade, $D \in \{0, 1\}$, has on metro journey times, $Y \subseteq \mathbb{R}$ (e.g. Canavan et al., 2015). In a two period setting, $t = (0, 1)$, the DGP before and after the upgraded signalling was introduced is given by

$$\begin{aligned} X_0 &\sim \mathcal{N}(\mu_X, \sigma_X^2) \\ D_1|X_0 &\sim \mathcal{B}(\text{expit}(\alpha X_0)) \\ Y_0|X_0, D_1 &\sim \mathcal{N}(\beta_{D_0} + D_1 + \beta_{X_0} X_0, \sigma_{Y_0}^2) \\ Y_1|X_0, D_1 &\sim \mathcal{N}(\beta_{D_1} + D_1 + \tau D_1 + \beta_{X_0} X_0, \sigma_{Y_1}^2) \\ (Y_1 - Y_0)|D_1 = \Delta Y_1|D_1 &\sim \mathcal{N}\left(\left(\beta_{D_1} - \beta_{D_0}\right) + \tau D_1, \sigma_{\Delta Y}\right) \end{aligned}$$

where we set $\tau = 4$ and $\beta_{D_1} - \beta_{D_0} = 2$. Note that time invariant confounders X_0 are removed via differencing and the parallel trend is satisfied.

We simulate two DID models.

1. $\widehat{\tau}_{DID1}$ - correctly specified DID regression model:

$$\mathbb{E}[Y|D] = \gamma_0 + \gamma_1 D_1 + \gamma_2 T + \tau_D (D_1 \times T) + \epsilon$$

Table 3
Simulation of DID estimator for CS3 ($\tau = 4.0$)

	Estimator	*Av. Est.*	*Emp. Var.*	*MSE*
1.	$\hat{\tau}_{DID1}$	4.000	0.008	0.008
2.	$\hat{\tau}_{DID2}$	3.000	0.010	1.001

Table 4
Simulation of RDD estimator for CS4 ($\tau = 5.0$)

	Estimator	*Av. Est.*	*Emp. Var.*	*MSE*
1.	$\hat{\tau}_{SRD1}$	4.995	0.015	0.015
2.	$\hat{\tau}_{SRD2}$	3.996	0.015	1.023
3.	$\hat{\tau}_{FRD}$	4.995	0.024	0.024

2. $\hat{\tau}_{DID2}$ - modified DGP in which the parallel trend is violated by introducing a covariate that influences only control in the post-treatment period, e.g.

$$Y_1|X_0, X_1, D_1 \sim \mathcal{N}(\beta_{D_1} + D_1 + \tau D_1 + \beta_{X_0} X_0 + \beta_{X_1} X_1 \times (1 - D_1), \sigma^2_{Y_1}).$$

Again the DID regression we use is: $\mathbb{E}[Y|D] = \gamma_0 + \gamma_1 D_1 + \gamma_2 T + \tau_D (D_1 \times T) + \epsilon$

The simulation results are shown in Table 3.

The correctly specified DID model produces a consistent estimate of τ when the parallel trend holds. Note, however, that the simulation demonstrates the necessity of this assumption, as illustrated in the bias of $\hat{\tau}_{DID2}$.

CS4: Environmental evaluation - simulation of RDD

Our objective is to quantify the causal effect that imposition of a low emission zone (LEZ), $D \in \{0, 1\}$, has on air quality $Y \subseteq \mathbb{R}$ (e.g. Ma et al., 2021). We have data before and after the imposition of the LEZ, and we use a sharp RDD model to evaluate its effect. The DGP for this case study is given by

$$T \sim \mathcal{U}(-1, 1)$$
$$Y \sim \mathcal{N}\left(\alpha + \beta T + \tau \times (T \geq 0), \sigma^2_Y\right)$$

where assignment to treatment, e.g. $D = 1[T \geq c]$, is determined by some threshold $c = 0$ of the forcing variable, which in this case could be time. The true ATE is given by parameter $\tau = 5.0$.

We simulate three RDD models.

1. $\hat{\tau}_{SRD1}$ - a correctly specified sharp RDD regression model:

$$Y = \alpha + \tau D + \beta_1 (T - c) + \beta_2 D(T - c) + \epsilon$$

2. $\hat{\tau}_{SRD2}$ - a modified DGP in which fuzziness is added to the threshold to simulate the presence of other factors that determine assignment to the treatment. Again, we estimate using a sharp RDD regression design, effectively ignoring the fact that treatment status is only partially determined by the running variable.
3. $\hat{\tau}_{FRD}$ - same DGP as RDD2, but this time using a fuzzy RDD estimator.

The simulation results are shown in Table 4.

Estimates of τ are unbiased under the sharp RDD estimator $SRD1$. When the DGP is modified to invalidate the assumption of continuity at the threshold, $SRD2$, the sharp RDD model fails to consistently estimate the true ATE. However, this can be rectified, as in $\hat{\tau}_{FRD}$, by dropping the assumption of sharp discontinuity in D and switching to a fuzzy RDD design.

8. Conclusions

This paper has provided an overview of the challenges we face in drawing causal inference in empirical transport research. It has outlined a framework for causal inference based on the concept of potential outcomes, and has discussed methods that can be used for estimation. Simulations have been presented to illustrate practical implementation of these methods in applied research.

The paper has shown that the models available to infer causal effects have their own particular inferential assumptions, which must be rigorously evaluated prior to application. Crucially, the issue of identification must feature as a core considerations in formulating our conceptual and methodological approaches to causal problems in transport research.

While the approaches discussed in this paper can offer transport researchers the prospect of obtaining a rigorous causal understanding of their data, several practical challenges remain that should be acknowledged.

First, is that for some methods (described in Section 4), all potential sources of confounding must be measured to obtain valid inference. This onerous requirement must be met to satisfy the so-called conditional independence assumption (CIA). In practice, we often face situations in which confounders are unobserved, or measured with error, leading to failure of the CIA. This problem is exacerbated by the fact that there are no reliable diagnostic procedures to comprehensively test for such failure.

Second, there are other methods (described in Section 5), which perhaps require less onerous data, but instead impose other stringent identifying assumptions. In the case of IV estimation, instruments must be relevant and exogenous, for DID models a parallel trend must hold, while continuity of outcomes and non-manipulation of the treatment threshold must hold for valid identification under RDD. It can be hard to establish empirically whether these assumptions hold in practice, and again diagnostics are not always clear cut.

Third, another major challenge for causal inference in our field relates to the requirement that the SUTVA is met, which applies to all methods covered in the paper. A key implication of the SUTVA is that the outcome for each unit must be independent of the treatment status of other units, or in other words, there should be no 'interference' in treatment effects between units. The assumption of no interference is often satisfied naturally when units are physically distinct and have no means of contact. But violations of the assumption can occur when proximity of units allows for contact, or when network-based interactions are present. This presents a particular concern for evaluation of causality in transport settings, in which local interventions can induce spatio-temporal propagations that affects other units throughout the entire network.

Appendix. Proofs

A.1. Identification of the average treatment effect (ATE) and average potential outcome (APO) under strong ignorability

A proof of theorem 2, in the case of binary treatments, is as follows.

Proof (*Identification of a Binary ATE Under Strong Ignorability*)**.**

$$\tau(1) = \mathbb{E}_i\left[Y_i(1) - Y_i(0)\right] \tag{28a}$$
$$= \mathbb{E}_X\left[\mathbb{E}_i(Y_i(1)|X_i = x) - \mathbb{E}_i(Y_i(0)|X_i = x)\right] \tag{28b}$$
$$= \mathbb{E}_X\left[\mathbb{E}_i(Y_i(1)|X_i = x, I_1(D_i) = 1) - \mathbb{E}_i(Y_i(0)|X_i = x, I_1(D_i) = 0)\right] \tag{28c}$$
$$= \mathbb{E}_X\left[\mathbb{E}_i(Y_i|X_i = x, I_1(D_i) = 1) - \mathbb{E}_i(Y_i|X_i = x, I_1(D_i) = 0)\right]. \tag{28d}$$

The law of iterated expectations gets us from ((28)a) to ((28)b), conditional independence justifies the equality of ((28)b) and ((28)c), the SUTVA allows the substitution of observed for potential outcomes to give ((28)d), and overlap ensures that the population ATE in ((28)d) is estimable since there are units in both the treated and untreated groups.

For continuous or multivalued treatments, identification of the APO, $\mu(d) = \mathbb{E}[Y_i(d)]$, under a given dose $D = d$, or the dose–response function, can be demonstrated as follows.

Proof (*Identification of a Multivalued or Continuous APO Under Strong Ignorability*)**.**

$$\mu(d) = \mathbb{E}[Y_i(d)] \tag{29a}$$
$$= \mathbb{E}_X\left[\mathbb{E}(Y_i(d)|X_i)\right] \tag{29b}$$
$$= \mathbb{E}_X\left[\mathbb{E}(Y_i(d)|I_d(D_i), X_i)\right] \tag{29c}$$
$$= \mathbb{E}_X\left[\mathbb{E}(Y_i|I_d(D_i), X_i)\right], \tag{29d}$$

where ((29)c) follows from conditional independence, ((29)d) from the SUTVA, and the overlap assumption ensures that ((29)d) is estimable since there are comparable units across treatment levels.

A.2. Balancing property of the propensity score

Lemma 1 (*Balancing of Pre-Treatment Covariates Given the Propensity Score*)**.** *If* $\pi\left(D_i|X_i\right)$ *is the propensity score, then*

$$I_d(D_i) \perp\!\!\!\perp X_i|\pi(d|X_i).$$

This follows because $\pi(d|X_i)$ *is a function of* X_i *and so conditioning on* X_i *adds no additional information*

$$\mathbb{E}\left[I_d(D_i)|X_i\right] = \mathbb{E}\left[I_d(D_i)|X_i, \pi(d|X_i)\right] = \mathbb{E}\left[I_d(D_i)|\pi(d|X_i)\right].$$

A proposition that follows immediately from this lemma is that conditional independence can be stated on the PS rather than X.

Corollary 1 (*Conditional independence given the propensity score*)**.** *Given balancing, we can establish conditional independence on the scalar PS rather than the potentially high-dimensional covariate vector* X_i*, e.g.*

$$Y_i(d) \perp\!\!\!\perp I_d(D_i)|\pi(d|X_i).$$

We can prove Lemma 1 in the case of binary and multivalued and continuous treatments respectively as follows.

Proof (*Balancing Property of the Propensity Score*)**.** First, in the case of a binary treatment

$$\Pr(D_i = 1|X_i, \pi(D_i = 1|X_i)) = \Pr(D_i = 1|X_i) = \pi(D_i = 1|X_i)$$

where the first equality follows from the fact that the PS is a function only of X_i and the second from the definition of the PS.

Furthermore, by the law of iterated expectation we have that,

$$\begin{aligned}
\Pr(D_i = 1|\pi(D_i = 1|X_i)) &= \mathbb{E}[I_1(D_i)|\pi(D_i = 1|X_i)] \\
&= \mathbb{E}_{\mathbb{X}}[\mathbb{E}\{I_1(D_i)|X_i, \pi(D_i = 1|X_i)\}|\pi(D_i = 1|X_i)] \\
&= \mathbb{E}_{\mathbb{X}}[\mathbb{E}\{I_1(D_i)|X_i\}|\pi(D_i = 1|X_i)] \\
&= \mathbb{E}[\Pr(D_i = 1|X_i)|\pi(D_i = 1|X_i)] \\
&= \mathbb{E}[\pi(D_i = 1|X_i)|\pi(D_i = 1|X_i)] \\
&= \pi(D_i = 1|X_i)
\end{aligned}$$

Hence we have that $\Pr(D_i = 1|X_i, \pi(D_i = 1|X_i)) = \mathbb{P}(D_i = 1|\pi(D_i = 1|X_i))$, and thus

$$I_1(D_i) \perp\!\!\!\perp X_i|\pi(D_i = 1|X_i).$$

For multivalued or continuous treatment we can apply the same logic. Let $\mathcal{X}$ be the sample space in which covariates X_i lie, then

$$\begin{aligned}
f_{D|\pi}\left(d|\pi(d|x_i)\right) &= \int_{\mathcal{X}} f_{D,X|\pi}\left(d, x_i|\pi(d|x_i)\right) dx_i \\
&= \int_{\mathcal{X}} f_{D|X,\pi}\left(d|x_i, \pi(d|x_i)\right) f_{X|\pi}\left(x_i|\pi(d|x_i)\right) dx_i \\
&= \int_{\mathcal{X}} f_{D|X}(d|x_i) f_{X|\pi}\left(x_i|\pi(d|x_i)\right) dx_i \\
&= \int_{\mathcal{X}} \pi(d|x_i) f_{X|\pi}\left(x_i|\pi(d|x_i)\right) dx_i, \\
&= \pi(d|x_i) = f_{D|X}(d|x_i).
\end{aligned}$$

Thus, $f_{D|\pi}\left(d|\pi(d|x_i)\right) = f_{D|X}(d|x_i)$, and therefore we have the property of balancing

$$\mathbb{E}\left[I_d(D_i)|X_i\right] = \mathbb{E}\left[I_d(D_i)|X_i, \pi(d|X_i)\right] = \mathbb{E}\left[I_d(D_i)|\pi(d|X_i)\right].$$

A.3. Identification via the propensity score

Below we prove Theorem 3 for multivalued or continuous treatments, with the analogous proof for a binary treatment just being the special case in which $d \in 0, 1$.

Proof (*Identification of an APO Under Strong Ignorability via PS Adjustment*)**.** For binary treatments

$$\mu(d) = \mathbb{E}\left[Y_i(d)\right] \tag{30a}$$
$$= \mathbb{E}_X\left[\mathbb{E}\left(Y_i(d)|X_i\right)\right] \tag{30b}$$
$$= \mathbb{E}_X\left[\mathbb{E}\left[Y_i(d)|\pi(d|X_i)\right]\right] \tag{30c}$$
$$= \mathbb{E}_X\left[\mathbb{E}\left[Y_i(d)|I_d(D_i), \pi(d|X_i)\right]\right] \tag{30d}$$
$$= \mathbb{E}_X\left[\mathbb{E}\left[Y_i|D_i = d, \pi(d|X_i)\right]\right]. \tag{30e}$$

A.4. Identification of an APO using inverse propensity score weighting

Proof (*Identification of an APO Using Inverse Propensity Score Weighting*)**.**

$$\begin{aligned}
\mathbb{E}_i\left[\frac{I_d(D_i) \cdot Y_i}{\pi(d|X_i; \alpha)}\right] &= \mathbb{E}_i\left[\frac{I_d(D_i) \cdot Y_i(d)}{\pi(d|X_i; \alpha)}\right] \\
&= \mathbb{E}_{X_i}\left[\mathbb{E}_i\left(\frac{I_d(D_i) \cdot Y_i(d)}{\pi(d|X_i; \alpha)}\middle|X_i\right)\right] \\
&= \mathbb{E}_{X_i}\left[\frac{\mathbb{E}_i(I_d(D_i)|X_i) \cdot \mathbb{E}_i(Y_i(d)|X_i)}{\pi(d|X_i; \alpha)}\right] \\
&= \mathbb{E}_{X_i}\left[\frac{\pi(d|X_i; \alpha) \cdot \mathbb{E}_i(Y_i(d)|X_i)}{\pi(d|X_i; \alpha)}\right] \\
&= \mathbb{E}_{X_i}[\mathbb{E}_i(Y_i(d)|X_i)]
\end{aligned}$$

$$= \mathbb{E}_i[Y_i(d)] = \mu(d)$$

where the first line uses SUTVA, the second uses iterated expectations, the third follows from conditional independence, and the fourth uses the propensity score definition.

A.5. Proof of the theorem of doubly robust estimation

The following proof of the DR property borrows from Naik et al. (2016) and is a direct extension of that given by Lunceford and Davidian (2004) from the binary treatment case to the multivalued case. As such the proof in the binary case follows as a special case.

Proof (*Double-Robustness*). For dose $d \in \mathcal{D}\{d_0, d_1, \ldots, d_m\}$ we write the DR estimator for $\hat{\mu}_{DR}(d)$ as

$$\hat{\mu}_{DR}(d) = \frac{1}{n}\sum_{i=1}^{n}\left[\Psi^{-1}\left\{m(d, X_i; \hat{\beta})\right\} + \frac{I_d(D_i)}{\hat{\pi}(d|X_i; \hat{\alpha})}\left[Y_i - \Psi^{-1}\left\{m(d, X_i; \hat{\beta})\right\}\right]\right]. \tag{31}$$

Applying the WLLN to (31) we have

$$\hat{\mu}_{DR}(d) \xrightarrow{p} \mathbb{E}[Y] + \mathbb{E}\left[\frac{I_d(D) - \pi(d|X;\alpha)}{\pi(d|X;\alpha)}\left\{Y - \Psi^{-1}\{m(d, X; \beta)\}\right\}\right], \tag{32}$$

which follows because

$$\mathbb{E}\left[\frac{I_d(D) - \pi(d|X;\alpha)}{\pi(d|X;\alpha)}Y\right] = \mathbb{E}\left[\frac{I_d(D)Y}{\pi(d|X;\alpha)}\right] - \mathbb{E}\left[\frac{\pi(d|X;\alpha)Y}{\pi(d|X;\alpha)}\right],$$

and where $\pi(d|X;\alpha)$ is the postulated PS model for $\pi(d|X)$, $\Psi^{-1}\{m(d, X; \beta)\}$ is the postulated OR model, and α and β are the "true" parameters of the PS and OR models respectively.

From the SUTVA, $Y(d) = I_d(D)Y$, thus (32) can be written

$$\mathbb{E}[Y(d)] + \mathbb{E}\left[\frac{I_d(D) - \pi(d|X;\alpha)}{\pi(d|X;\alpha)}\left\{Y(d) - \Psi^{-1}\{m(d, X; \beta)\}\right\}\right], \tag{33}$$

The first term of (33) is what we are trying to estimate, so all that remains is to show that the second term is 0 if either the PS or OR model is correctly specified. Let us first consider the case where the OR model is correctly specified, so that

$$\Psi^{-1}\{m(d, X; \beta)\} = \mathbb{E}(Y|D = d, X)$$

Then, using the above equality and the law of iterated expectation, the second term of (3.4) becomes

$$\begin{aligned}
&\mathbb{E}\left(\mathbb{E}\left[\frac{I_d(D) - \pi(d|X;\alpha)}{\pi(d|X;\alpha)}\left\{Y(d) - \mathbb{E}(Y|D = d, X)\right\}\middle| I_d(D), X\right]\right)\\
=&\mathbb{E}\left(\frac{I_d(D) - \pi(d|X;\alpha)}{\pi(d|X;\alpha)}\mathbb{E}\left[\left\{Y(d) - \mathbb{E}(Y|D = d, X)\right\}\middle| I_d(D), X\right]\right)\\
=&\mathbb{E}\left(\frac{I_d(D) - \pi(d|X;\alpha)}{\pi(d|X;\alpha)}\left\{\mathbb{E}[Y(d)|I_d(D), X] - \mathbb{E}(Y|D = d, X)\right\}\right)\\
=&\mathbb{E}\left(\frac{I_d(D) - \pi(d|X;\alpha)}{\pi(d|X;\alpha)}\left\{\mathbb{E}[Y(d)|X] - \mathbb{E}(Y(d)|X)\right\}\right) = 0
\end{aligned}$$

where the last line follows from the weak conditional independence condition implied by the no unmeasured confounders assumption, which gives that

$$\mathbb{E}(Y|D = d, X) = \mathbb{E}(Y(d)|D = d, X) = \mathbb{E}(Y(d)|X) = \mathbb{E}(Y(d)|I_d(D), X)$$

Conversely, if the PS model is correctly specified then

$$\pi(d|X;\alpha) = \pi(d|X) = \mathbb{P}(D = d|X) = \mathbb{E}(I_d(D)|X)$$

Thus, using the above equality and the law of iterated expectation, the second term of (3.4) becomes

$$\begin{aligned}
&\mathbb{E}\left(\mathbb{E}\left[\frac{I_d(D) - \pi(d|X)}{\pi(d|X)}\left\{Y(d) - \Psi^{-1}\{m(d, X; \beta)\}\right\}\middle| Y(d), X\right]\right)\\
=&\mathbb{E}\left(\left\{Y(d) - \Psi^{-1}\{m(d, X; \beta)\}\right\}\mathbb{E}\left[\frac{I_d(D) - \pi(d|X)}{\pi(d|X)}\middle| Y(d), X\right]\right)\\
=&\mathbb{E}\left(\left\{Y(d) - \Psi^{-1}\{m(d, X; \beta)\}\right\}\frac{\mathbb{E}[I_d(D)|Y(d), X] - \pi(d|X)}{\pi(d|X)}\right)\\
=&\mathbb{E}\left(\left\{Y(d) - \Psi^{-1}\{m(d, X; \beta)\}\right\}\frac{\mathbb{E}[I_d(D)|X] - \pi(d|X)}{\pi(d|X)}\right)\\
=&\mathbb{E}\left(\left\{Y(d) - \Psi^{-1}\{m(d, X; \beta)\}\right\}\frac{\pi(d|X) - \pi(d|X)}{\pi(d|X)}\right) = 0
\end{aligned}$$

where the second last line follows from the no unmeasured confounders assumption as before.

Thus we see that $\hat{\mu}_{DR}(d)$ consistently estimates the APO $\mu(d)$ when either the OR model or PS model is correctly specified, in the case of multivalued treatment.

A.6. Identification via instrumental variables

We have defined the IV estimator for β_D as

$$\beta_{D_{IV}} = \mathbb{E}[(W^\top D)]^{-1}\mathbb{E}[(W^\top Y)].$$

This quantity is identified if the expectations $\mathbb{E}[(W^\top D)]$ and $\mathbb{E}[(W^\top Y)]$ can be consistently estimated using observed data.

We can demonstrate consistency under the IV assumptions as follows. If W is uncorrelated with e such that plim $(n^{-1}W^\top e) = 0$, and, if W is associated with D such that plim $(n^{-1}W^\top D) = \Sigma_{WD}$ exists and is non-singular, then

$$\begin{aligned} W^\top Y &= W^\top D\beta_D + W^\top e \\ n^{-1}W^\top Y &= n^{-1}W^\top D\beta_D + n^{-1}W^\top e \\ \text{plim}(n^{-1}W^\top Y) &= \text{plim}(n^{-1}W^\top D)\beta_D + \text{plim}(n^{-1}W^\top e) \end{aligned}$$

Since $n^{-1}W^\top e \xrightarrow{p} \mathbb{E}[W^\top e] = 0$ as $n \to \infty$, then using sample moments to consistently estimate $\mathbb{E}[(W^\top Y)]$ and $\mathbb{E}[(W^\top D)]$, we have

$$\hat{\beta_{DIV}} = \left(n^{-1}W^\top D\right)^{-1} n^{-1}W^\top Y = \frac{\text{Cov}(W,Y)}{\text{Cov}(W,D)}.$$

References

Abadie, A., 2021. Using synthetic controls: Feasibility, data requirements, and methodological aspects. J. Econ. Lit. 59 (2), 391–425.

Abadie, A., Diamond, A., Hainmueller, J., 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. J. Amer. Statist. Assoc. 105 (490), 493–505.

Abadie, A., Gardeazabal, J., 2003. The economic costs of conflict: a case study of the basque country. Amer. Econ. Rev. 93 (1), 112–132.

Abadie, A., Imbens, G., 2002. Simple and Bias-Corrected Matching Estimators for Average Treatment Effects. Technical Working Paper 283, National Bureau of Economic Research.

Abadie, A., Imbens, G.W., 2006. Large sample properties of matching estimators for average treatment effects. Econometrica 74 (1), 235–267.

Athey, S., Imbens, G.W., 2017. The state of applied econometrics: Causality and policy evaluation. J. Econ. Perspect. 31 (2), 3–32.

Bang, H., Robins, J.M., 2005. Doubly robust estimation in missing data and causal inference models. Biometrics 61, 962–972.

Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust difference-in-differences estimates? Q. J. Econ. 119 (1), 249–275.

Canavan, S., Graham, D.J., Melo, P.C., Anderson, R.J., Barron, A.S., Cohen, J.M., 2015. Impacts of moving-block signaling on technical efficiency application of propensity score matching on urban metro rail systems. Transp. Res. Rec. 68–74.

Cox, D.R., 1958. Planning of Experiments. John Wiley & Sons, London.

Fisher, R.A., 1935. The Design of Experiments. Oliver and Boyd, Edinburgh.

Graham, D.J., 2021. Causal inference for ex post evaluation of transport interventions. In: Vickerman, R. (Ed.), International Encyclopedia of Transportation. Elsevier, Oxford, pp. 283–290.

Graham, D.J., McCoy, E.J., Stephens, D.A., 2013. Quantifying the effect of area deprivation on child pedestrian casualties using longitudinal mixed models to adjust for confounding, interference, and spatial dependence. J. Roy. Statist. Soc. Ser. A 176 (4), 931–950.

Graham, D.J., McCoy, E.J., Stephens, D.A., 2014. Quantifying causal effects of road network capacity expansions on traffic volume and density via a mixed model propensity score estimator. J. Amer. Statist. Assoc. 109 (508), 1440–1449.

Graham, D.J., McCoy, E.J., Stephens, D.A., 2016. Approximate Bayesian inference for doubly robust estimation. Bayesian Anal. 11 (1), 47–69.

Graham, D.J., Naik, C., McCoy, E.J., Li, H., 2019. Do speed cameras reduce road traffic collisions? PLoS One 14.

Hahn, J., Hausman, J., 2003. Weak instruments: diagnosis and cures in empirical economics. Amer. Econ. Rev. 93, 118–125.

Handy, S., Cao, X., Mokhtarian, P., 2005. Correlation or causality between the built environment and travel behavior? Evidence from Northern California. Transp. Res. D 10 (6), 427–444.

Hirano, K., Imbens, G.W., 2004. The propensity score with continuous treatments. In: Gelman, A., Meng, X.L. (Eds.), Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives. Wiley, New York, pp. 73–84.

Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrics 71 (4), 1161–1189.

Holland, P.W., 1986. Statistics and causal inference. J. Amer. Statist. Assoc. 81 (396), 945–970.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. 47, 663–685.

Imbens, G.W., 2000. The role of the propensity score in estimating dose-response functions. Biometrika 87 (3), 706–710.

Imbens, G.W., 2014. Instrumental variables: An econometrician's perspective. Statist. Sci. 29 (3), 323–358.

Imbens, G., Lemieux, T., 2008. Regression discontinuity designs: A guide to practice. J. Econometrics 142, 613–635.

Imbens, G.W., Rubin, D.B., 2015. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, New York, NY, USA.

Imbens, G.W., Wooldridge, J.M., 2009. Recent developments in the econometrics of program evaluation. J. Econ. Lit. 47 (1), 5–86.

Joffe, M.M., Rosenbaum, P.R., 1999. Propensity scores. Am. J. Epidemiol. 150 (4), 327–333.

Kang, J.D.Y., Schafer, J.L., 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statist. Sci. 22 (4), 523–539.

van der Laan, M., Robins, J.M., 2003. Unified Methods for Censored Longitudinal Data and Causality. Springer, Berlin.

Lee, D.S., Lemieux, T., 2010. Regression discontinuity designs in economics. J. Econometrics 48, 281–355.

Li, H., Graham, D.J., 2016. Heterogeneous treatment effects of speed cameras on road safety. Accid. Anal. Prev. 97, 153–161.

Lunceford, J.K., Davidian, M., 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat. Med. 23, 2937–2960.

Ma, L., Graham, D.J., Stettler, M.E.J., 2021. Has the ultra low emission zone in London improved air quality? Environ. Res. Lett. 16 (12), 124001.

Naik, C., McCoy, E.J., Graham, D.J., 2016. Multiply robust dose-response estimation for multivalued causal inference problems. ArXiv.

Neyman, J., 1923. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statist. Sci. 5 (4), 465–480, translated in 1990.
Robins, J.M., 1986. A new approach to causal inference in mortality studies with a sustained exposure period: application to control for the healthy worker survivor effect. Math. Model. 7, 1393–1512.
Robins, J.M., 1999a. Association, causation, and marginal structural models. Synthese 121, 151–179.
Robins, J.M., 1999b. Marginal structural models versus structural nested models as tools for causal inference. In: Statistical Models in Epidemiology: The Environmental and Clinical Trials. Springer-Verlag, New York, pp. 95–134.
Robins, J.M., 2000. Robust estimation in sequentially ignorable missing data and causal inference models. In: Proceedings of the American Statistical Association, Section on Bayesian Statistical Science. American Statistical Association, Alexandria, VA, pp. 6–10.
Robins, J.M., Hernán, M.A., Brumback, B., 2000a. Marginal structural models and causal inference in epidemiology. Epidemiology 11 (5), 550–560.
Robins, J.M., Rotnitzky, A., 2001. Comment on "Inference for semiparametric models: some questions and an answer". Stat. Sin. 11, 920–936.
Robins, J.M., Rotnitzky, A., van der Laan, M.J., 2000b. Comment on the Murphy and Van der Vaart article "On profile likelihood". J. Amer. Statist. Assoc. 95, 431–435.
Rosenbaum, P.R., 1987. Model-based direct adjustment. J. Amer. Statist. Assoc. 82 (398), 387–394.
Rosenbaum, P.R., 1989. Optimal matching for observational studies. J. Amer. Statist. Assoc. 84 (408), 1024–1032.
Rosenbaum, P.R., 1999. Propensity score. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics. John Wiley, New York, pp. 3551–3555.
Rosenbaum, P.R., 2002. Covariance adjustment in randomized experiments and observational studies. Statist. Sci. 17 (3), 286–304.
Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.
Rubin, D.B., 1973a. Matching to remove bias in observational studies. Biometrics 29, 159–183.
Rubin, D.B., 1973b. The use of matched sampling and regression adjustments to remove bias in observational studies. Biometrics 29, 185–203.
Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and non-randomized studies. J. Educ. Psychol. 66 (5), 688–701.
Rubin, D.B., 1977. Assignment to treatment group on the basis of a covariate. J. Educ. Stat. 2 (1), 1–26.
Rubin, D.B., 1978. Bayesian inference for causal effects: the role of randomization. Ann. Statist. 6 (1), 34–58.
Rubin, D.B., 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. J. Amer. Statist. Assoc. 74 (366), 318–328.
Rubin, D.B., 1980. Comment on 'Randomization analysis of experimental data in the Fisher randomization test' by Basu. J. Amer. Statist. Assoc. 75 (371), 591–593.
Rubin, D.B., 1986. Comment: which ifs have causal answers? J. Amer. Statist. Assoc. 81 (396), 961–962.
Rubin, D.B., 1990. Neyman (1923) and causal inference in experiments and observational studies. Statist. Sci. 5 (4), 472–480.
Rubin, D.B., 2006. Using propensity scores to help design observational studies: applications to the tobacco litigation. In: Hantula, D. (Ed.), Advances in Social and Organizational Psychology. Erlbaum, Mahwah, NJ, pp. 41–59.
Rubin, D.B., Thomas, N., 1996. Matching using estimated propensity scores: relating theory to practice. Biometrics 52 (1), 249–264.
Rubin, D.B., Thomas, N., 2000. Combining propensity score matching with additional adjustments for prognostic covariates. J. Amer. Statist. Assoc. 95 (450), 573–585.
Scharfstein, D.O., Rotnitzky, A., Robins, J.M., 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. J. Amer. Statist. Assoc. 94 (448), 1096–1120, (with rejoinder 1135-1146).
Tsiatis, A.A., 2006. Semiparametric Theory and Missing Data. Springer, Berlin.
Tsiatis, A.A., Davidian, M., 2007. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statist. Sci. 22 (4), 569–573.
Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data, second ed. MIT Press, Cambridge, Massachusetts.