The causalweight Package

Hugo Bodory
University of Fribourg

Martin Huber
University of Fribourg

Abstract

We describe the R package **causalweight** for causal inference based on inverse probability weighting (IPW). The **causalweight** package offers a range of semiparametric methods for treatment or impact evaluation and mediation analysis, which incorporates intermediate outcomes for investigating causal mechanisms. Depending on the method, identification relies on selection on observables assumptions or on instrumental variables when selection is on unobservables, approaches that may also be applied to tackle non-random outcome attrition and sample selection. Inference is based on the bootstrap.

Keywords: Treatment effect, selection on observables, sample selection, mediation analysis, instrumental variable, IPW.

1. Introduction

Researchers in epidemiology, economics, political sciences, or other social sciences frequently aim at evaluating the causal effect of some binary intervention or treatment, as well as learning about the mechanisms through which a causal effect operates. This paper introduces the R package **causalweight** for analyzing the causal effect of a binary treatment as well as its mechanisms (based on mediation analysis that incorporates intermediate outcomes called mediators) under various identifying assumptions. All estimators rely on some form of inverse probability weighting (IPW), by weighing outcomes by the inverse of a specific conditional probability or propensity score. The **causalweight** package includes treatment evaluation under treatment selection on observables with and without controlling for non-random outcome attrition or sample selection (Huber 2012, 2014b), instrumental variable-based estimation of local average treatment effects when controlling for observed covariates (Frölich 2007), and mediation analysis for investigating causal mechanisms with selection on observables or instrumental variable assumptions (Huber 2014a; Frölich and Huber 2017). The nonparametric identification strategies underlying the estimators avoid imposing strong functional form restrictions in the structural models considered. Estimation of the propensity scores relies on probit or logit specifications.

In the next chapters, we discuss various treatment effect models along with methods for analysing them and demonstrate the functionalities of the R package **causalweight** by means of examples with simulated data. Section 2 presents an overview of the functions available in the **causalweight** package. Section 3 discusses a treatment effect model with treatment selection on observables and non-random outcome attrition or sample selection. It also introduces the function `treatweight`, which allows treatment evaluation with and without sample selection correction, either based on observables or an instrument for selection. Section 4 presents causal

mediation models based on selection on observables assumptions along with the `medweight` function for estimating causal mechanisms. Section 5 discusses treatment effect evaluation based on an instrument when controlling for observed covariates and its implementation in the `lateweight` function. Section 6 considers mediation analysis with distinct instruments for the treatment and the mediator when controlling for observed covariates, as implemented in the `medlateweight` function. Section 7 concludes.

2. Overview of the causalweight package

The **causalweight** package consists of four functions aimed at user-friendly treatment evaluation and mediation analysis. The following table illustrates the structure of the **causalweight** package by assigning to each of the main functions the corresponding treatment effect/mediation model.

Table 1: Main functions of the **causalweight** package

Functions in R	Treatment effect models
`treatweight`	Treatment evaluation with or without sample selection correction (Section 3).
`medweight`	Causal mediation analysis (Section 4).
`lateweight`	Local average treatment effect with covariates (Section 5).
`medlateweight`	Causal mediation analysis with instrumental variables (Section 6).

The function `treatweight` implements treatment evaluation under treatment selection on observables, optionally with correcting for sample selection or non-ignorable outcome attrition based on either a selection on observables/missing at random assumption or an instrument. To tackle the double selection problem into the treatment and into the subpopulation with non-missing outcomes, it makes use of both treatment and selection propensity scores to appropriately reweigh observations by IPW, see Huber (2012, 2014b). The function `treatweight` allows computing the average treatment effect in the total population (ATE) and on the treated (ATET).

The function `medweight` implements mediation analysis to investigate the causal mechanisms of a binary treatment under selection on observables based on IPW. More specifically, it computes (i) the (total) average treatment effect, (ii) the average natural *indirect* effect, which operates through an intermediate outcome (or mediator) situated on the causal path between the treatment and the outcome, and (iii) the (unmediated) average natural *direct* effect, see Huber (2014a). The *indirect* and *direct* effect estimates are returned under either potential treatment state. The function `treatweight` allows computing the effects for both the total population and the subpopulation of the treated.

The function `lateweight` returns the local average treatment effect (LATE) of a binary endogenous treatment based on IPW using a binary endogenous instrument that is conditionally valid given observed covariates, see Frölich (2007). In addition, it returns the intention-to-treat effect of the instrument on the outcome, as well as first-stage effect of the instrument on the treatment. The function `lateweight` permits estimating the local average treatment effect among all subjects whose treatment complies with the instrument (LATE) and among treated compliers (LATTs) by weighing units by the inverse of their instrument propensity scores.

The function `medlateweight` computes the causal mechanisms (natural direct and indirect effects) of a binary treatment among treatment compliers based on distinct instrumental variables (IVs) for the treatment and the mediator, which are assumed to be conditionally valid given a set of observed covariates. The treatment and its instrument are assumed to be binary while the mediator and its instrument are assumed to be continuous. This motivates combining the LATE approach with a control function approach for tackling mediator endogeneity, see Theorem 1 in Frölich and Huber (2017). The function `medlateweight` yields (i) the (total) local average treatment effect (LATE) among compliers based on IPW, (ii) the average natural *direct* and *indirect* effects under either potential treatment state among compliers based on IPW, and (iii) parametric direct and indirect effect estimates (imposing effect homogeneity across treatment states) based on regression.

Details on the models and the implementation of the corresponding estimators in the **causalweight** package are provided in the following Sections 3 to 6.

3. Treatment evaluation with sample selection correction

The function `treatweight` implements treatment effect evaluation when the treatment selection is related to observed covariates, optionally with considering sample selection/non-random outcome attrition. The latter case constitutes a double selection problem (i) into the treatment (selection on observables) and (ii) into the subpopulation for which the outcome is observed (selection on unobservables). The function `treatweight` computes the average treatment effect (ATE) and the average treatment effect on the treated (ATET) by weighing observations by the inverse of (nested) propensity scores. The nested weights control for treatment selection bias due to non-random treatment assignment and sample selection bias in the subpopulation with observed outcomes, see Huber (2012, 2014b).

3.1. Model

When estimating the causal effect of a binary treatment D on an outcome Y, researchers are typically confronted with the identification issue that take-up of D is selective. As a further complication, Y might only be observed for a subpopulation that is non-randomly selected, as indicated by a binary sample selection variable S. We tackle the former issue by assuming treatment selection on observed covariates X and the latter issue by either assuming ignorability of sample selection given observables, or the availability of an instrument Z that is conditionally valid.

We consider a general model, in which outcome Y is an unknown function of two observed components, the binary treatment D and the vector of covariates X, and a possibly multidimensional unobserved term U:

$$Y = \varphi(D, X, U), \tag{1}$$

where $\varphi(\cdot)$ is an unknown function.

While D and X are assumed to be observed for everyone, the `treatweight` function permit for sample selection, implying that outcome Y is only observed for a subpopulation as indicated by the binary selection indicator S. Empirical examples for such set-ups include wage equations (where S is employment), see Gronau (1974) and Heckman (1976, 1974), the evaluation of

effects of educational interventions on test scores (where S is participation in the test), see Angrist, Bettinger, and Kremer (2006) and Angrist, Lang, and Oreopoulos (2009), or loss of outcome follow-up in repeated surveys. In our model, the selection indicator is either assumed to be a function of the treatment, the covariates, and an unobserved term, or of the previously mentioned terms and an instrument:

$$S = I\{\eta(D,X) \geq V\} \text{ (scenario 1)}, \quad (2)$$
$$S = I\{\zeta(D,X,Z) \geq V\} \text{ (scenario 2)}. \quad (3)$$

$I\{\cdot\}$ denotes the indicator function and $\eta(\cdot), \zeta(\cdot)$ are unknown functions. Z represents a one or multi-dimensional instrument which is observable for all units and not directly related with the outcome. V is an unobserved term. If it is not associated with U, sample selection is related to observables or missing at random (MAR) in the denomination of Rubin (1976). If V is associated with U, S is endogenous even when controlling for (D,X). In this case, identification crucially hinges on the availability of an instrument Z that is relevant for S in the sense that it shifts the selection probability conditional on (D,X) but does not appear in φ (exclusion restriction), as in scenario 2 of (2). In general, at least one element in Z needs to be continuous.

To define the causal effect of D, we utilize the potential outcome framework advocated by Rubin (1974), among others. We denote the potential outcome for individual i and some hypothetical treatment $D = d$ as

$$Y_i(d) = \varphi(d, X_i, U_i). \quad (4)$$

The difference $Y_i(1) - Y_i(0)$ would yield the individual treatment effect, but is unknown to the researcher, because each individual is either treated or not treated and cannot appear in both states of the world at the same time. The average treatment effect (ATE), which can be identified under assumptions outlined in the following section, is given by the mean difference of the potential outcomes under treatment and non-treatment:

$$\Delta = \mathsf{E}[Y(1)] - \mathsf{E}[Y(0)]. \quad (5)$$

A further parameter of policy interest, is the mean effect among those receiving the treatment, the average treatment effect on the treated (ATET):

$$\Delta_{D=1} = \mathsf{E}[Y(1)|D=1] - \mathsf{E}[Y(0)|D=1]. \quad (6)$$

3.2. Identification

In the absence of sample selection, the ATE is identified if $Y(1), Y(0)$ are independent of D conditional on X (selection on observables) and the treatment propensity score $\pi(X) \equiv \Pr(D=1|X)$ is larger than zero and smaller than 1 almost surely (common support), see Imbens and Wooldridge (2009). The ATE then corresponds to the following expression based on weighing by the inverse of the propensity score:

$$\Delta = \mathsf{E}\left[\frac{D \cdot Y}{\pi(X)}\right] - \mathsf{E}\left[\frac{(1-D) \cdot Y}{1-\pi(X)}\right]. \quad (7)$$

The idea of inverse probability weighting (IPW) goes back to Horvitz and Thompson (1952), who first proposed an estimator of the population mean in the presence of non-randomly missing data. The ATET is obtained by multiplying the expressions in the expectation operators of (7) by $\pi(X)/\Pr(D=1)$, see Hirano, Imbens, and Ridder (2003), which corresponds to:

$$\Delta_{D=1} = \mathsf{E}\left[\frac{D\cdot Y}{\Pr(D=1)}\right] - \mathsf{E}\left[\frac{(1-D)\cdot Y\cdot \pi(X)}{(1-\pi(X))\cdot \Pr(D=1)}\right]. \tag{8}$$

The ATET is identified under the assumptions that $Y(0)$ is independent of D conditional on X and $\pi(X)$ is smaller than 1 almost surely.

Complications prevail if the outcomes are observed for a selective subpopulation only, which requires further assumptions for identification. One possible condition is that sample selection S is driven by observable variables but independent of Y conditional on (D,X), i.e. MAR (see scenario 1 in (2)). When adding this assumption to the previous ones, the ATE is identified by reweighing observations (additionally to the inverse treatment propensity score) by the inverse of the sample selection propensity score $p(D,X)\equiv\Pr(S=1|D,X)$, see Huber (2012):

$$\Delta = \mathsf{E}\left[\frac{S\cdot D\cdot Y}{p(D,X)\cdot \pi(X)}\right] - \mathsf{E}\left[\frac{S\cdot (1-D)\cdot Y}{p(D,X)\cdot (1-\pi(X))}\right], \tag{9}$$

which hinges on $p(D,X)$ being larger than 0 almost everywhere as additional common support restriction.

Alternatively to assuming MAR, sample selection might be tackled by an instrumental variable strategy, see see scenario 2 in (2). In this context, Δ is identified under the following assumptions, see Huber (2014b): (i) satisfaction of the selection on observables assumption in the total population as before, (ii) availability of an instrument for selection that satisfies the exclusion restriction such that the sample selection propensity score $\Pr(S=1|D,X,Z)$ is a valid control function, (iii) independence of (U,V) and (D,Z) conditional on $\Pr(S=1|D,X,Z)$ and X, and (iv) homogeneity of average treatment effects conditional on X. The ATE on the total population under sample selection is identified by weighing by the inverse of a nested treatment propensity score as well as the selection propensity score, given that specific common support conditions on the propensity scores hold:

$$\Delta = \mathsf{E}\left[\frac{S\cdot D\cdot Y}{p(W)\cdot \pi(X,p(W))}\right] - \mathsf{E}\left[\frac{S\cdot (1-D)\cdot Y}{p(W)\cdot (1-\pi(X,p(W)))}\right]. \tag{10}$$

$\pi(X,p(W))$ denotes the treatment propensity score $\Pr(D=1|X,p(W))$, i.e., the probability of being treated conditional on X and $p(W)$, with $W\equiv(D,X,Z)$ and $p(W)\equiv\Pr(S=1|D,X,Z)$. Analogously to (8), multiplying the expressions in the expectation operators of (9) and (10) by $\pi(X)/\Pr(D=1)$ yields the ATET under the respective set of assumptions.

3.3. Estimation

Assuming an i.i.d. sample of n units prior to selection, indexed by $i=1,...,n$, we briefly discuss the estimation of the ATE under sample selection using an instrument based on the normalized sample analog of (10). Estimation of treatment effects under different sets of assumptions (i.e. MAR or no sample selection) proceeds analogously. Let $\hat{p}(W)$ and $\hat{\pi}(X,\hat{p}(W))$ denote estimates of the sample selection propensity score $p(W)$ and the treatment propensity

score $\pi(X, p(W))$, respectively. A general 3-step estimation approach proceeds as follows:

(a) Estimate $\hat{p}(W)$ by regressing S on $(1, D, X, Z)$,
(b) estimate $\hat{\pi}(X, \hat{p}(W))$ by regressing D on $(1, X, \hat{p}(W))$,
(c) obtain an estimate of the ATE, denoted by $\hat{\Delta}$, as the normalized sample analogue of (10) in which $\hat{p}(W)$ and $\hat{\pi}(X, \hat{p}(W))$ are used as plug-in estimates.

The propensity scores are estimated by probit or logit models. The normalized sample analogue of (10) corresponds to

$$\begin{aligned}\hat{\Delta} &= \sum_{i=1}^{n} \frac{S_i \cdot D_i \cdot Y_i}{\hat{\pi}(X_i, \hat{p}(W_i))} \Big/ \sum_{i=1}^{n} \frac{S_i \cdot D_i}{\hat{\pi}(X_i, \hat{p}(W_i))} \\ &- \sum_{i=1}^{n} \frac{S_i \cdot (1 - D_i) \cdot Y_i}{\hat{p}(W_j) \cdot (1 - \hat{\pi}(X_i, \hat{p}(W_i)))} \Big/ \sum_{i=1}^{n} \frac{S_i \cdot (1 - D_i)}{\hat{p}(W_j) \cdot (1 - \hat{\pi}(X_i, \hat{p}(W_i)))}. \end{aligned} \tag{11}$$

The normalizations $\sum_{i=1}^{n} \frac{S_i \cdot D_i}{\hat{\pi}(X_i, \hat{p}(W_i))}$ and $\sum_{i=1}^{n} \frac{S_i \cdot (1-D_i)}{\hat{p}(W_j) \cdot (1-\hat{\pi}(X_i, \hat{p}(W_i)))}$ ensure that the weights in each treatment group add up to unity. This may improve the finite sample properties of the estimator, see for instance the discussions in Imbens (2004) and Busso, DiNardo, and McCrary (2014).

This and other semiparametric IPW estimators discussed further below can be expressed as sequential GMM estimators where parametric propensity score estimation represents the first step and effect estimation the second step, see Newey (1984). It follows from his results that our methods are $\sqrt{n}$-consistent and asymptotically normal under standard regularity conditions. Therefore, the i.i.d. bootstrap is a valid inference method for treatment effect estimators based on IPW, see also Hirano *et al.* (2003). The function `treatweight` allows specifying the number of bootstrap replications for computing standard errors. Furthermore, it offers a trimming rule for discarding observations with extreme propensity scores to improve overlap, see Crump, Hotz, Imbens, and Mitnik (2009). The default is to discard observations with treatment propensity scores smaller than 0.05 (5%) or larger than 0.95 (95%), when considering the ATE or larger than 0.95 when considering the ATET. When a sample selection correction is included, the default is to discard observations with sample selection propensity scores smaller than 0.05.

3.4. Examples in R

This section presents (i) the input arguments of the `treatweight` function, (ii) the output stored in the object generated by `treatweight`, and (iii) two examples for ATE estimation with and without sample selection, respectively.

Input arguments of `treatweight`

The input arguments of `treatweight` are:

Table 2: Input arguments of the `treatweight` function

Variables	Features of the variables
`y`	Dependent variable.

continued ...

... continued

Variables	Features of the variables
`d`	Treatment, must be binary (either 1 or 0), must not contain missings.
`x`	Confounders of the treatment and outcome, must not contain missings.
`s`	Selection indicator. Must be 1 if `y` is observed (non-missing) and 0 if `y` is not observed (missing). Default is `NULL`, implying that `y` does not contain any missings.
`z`	Optional instrumental variable(s) for selection `s`. If `NULL`, outcome selection based on observables (`x`,`d`) - known as "missing at random" - is assumed. If `z` is defined, outcome selection based on unobservables - known as "non-ignorable missingness" - is assumed. Default is `NULL`. If `s` is `NULL`, `z` is ignored.
`selpop`	Only to be used if both `s` and `z` are defined. If `TRUE`, the effect is estimated for the selected subpopulation with $s = 1$ only. If FALSE, the effect is estimated for the total population (note that this relies on somewhat stronger statistical assumptions). Default is `FALSE`. If `s` or `z` is `NULL`, `selpop` is ignored.
`ATET`	If `FALSE`, the average treatment effect (`ATE`) is estimated. If `TRUE`, the average treatment effect on the treated (`ATET`) is estimated. Default is `FALSE`.
`trim`	Trimming rule for discarding observations with extreme propensity scores. If `ATET` = `FALSE`, observations with $\Pr(D = 1\|X) <$ `trim` or $\Pr(D = 1\|X) >$ (1−`trim`) are dropped. If `ATET` = `TRUE`, only those observations with $\Pr(D = 1\|X) >$ (1−`trim`) are dropped. If `s` is defined and `z` is `NULL`, observations with extremely low selection propensity scores, $\Pr(S = 1\|D, X) <$ `trim`, are discarded, too. If `s` and `z` are defined, the treatment propensity scores to be trimmed change to $\Pr(D = 1\|X, \Pr(S = 1\|D, X, Z))$. If in addition `selpop` = `TRUE`, observation with $\Pr(S = 1\|D, X, Z) <$ `trim` are discarded, too. Default for `trim` is 0.05.
`logit`	If `FALSE`, probit regression is used for propensity score estimation. If `TRUE`, logit regression is used. Default is `FALSE`.
`boot`	Number of bootstrap replications for estimating standard errors. Default is 1999.

The `treatweight` *object*

A `treatweight` object contains six components all of which can be referenced by a dollar sign (`$`), see the examples in this section below. These components are:

Table 3: Components of the `treatweight` object

Components	Description of the components
`effect`	Average treatment effect (`ATE`) if `ATET` = `FALSE` or the average treatment effect on the treated (`ATET`) if `ATET` = `TRUE`.
`se`	bootstrap-based standard error of the effect.
`pval`	p-value of the effect.

continued ...

... continued

Components	Description of the components
y1	mean potential outcome under treatment.
y0	mean potential outcome under control.
ntrimmed	number of discarded (trimmed) observations due to extreme propensity score values.

Example for estimating the ATE without sample selection

This example estimates the ATE based on equation (7) in simulated data. The sample size n is set to 10'000. The seeds set when generating random variables (set.seed()) enable the replication of the results. The following chunk of R input code results in the output of the function treatweight:

```
> n=10000
> set.seed(100); x=rnorm(n)
> set.seed(101); d=(0.25*x+rnorm(n)>0)*1
> set.seed(102); y=0.5*d+0.25*x+rnorm(n)
> output=treatweight(y=y,d=d,x=x,trim=0.05,ATET=FALSE,logit=TRUE, boot=19)
> cat("ATE: ",round(c(output$effect),3),", standard error: ",
+     round(c(output$se),3), ", p-value: ",round(c(output$pval),3))
> output$ntrimmed
```

The following chunk of output code displays two lines (based on the treatweight object called output). The first line gives the ATE estimate, standard error, and p-value, respectively (rounded to three decimals). The second line provides the number of observations discarded by the trimming rule.

```
ATE:  0.488 , standard error:  0.022 , p-value:  0

[1] 0
```

Example for estimating the ATE under sample selection based on an instrument

This example estimates the ATE under sample selection based on equation (10) in simulated data. The sample size n is set to 10'000. Matrix e reflects the unobserved terms of equations (1) and (2) for computing y and s and follows a multivariate normal distribution with covariance matrix sigma. The following chunk of R input code results in the output of the function treatweight:

```
> n=10000
> sigma=matrix(c(1,0.6,0.6,1),2,2)
> set.seed(100); e=(2*rmvnorm(n,rep(0,2),sigma))
> set.seed(101); x=rnorm(n)
> set.seed(102); d=(0.5*x+rnorm(n)>0)*1
```

```
> set.seed(103); z=rnorm(n)
> s=(0.25*x+0.25*d+0.5*z+e[,1]>0)*1
> y=d+x+e[,2]; y[s==0]=0
> output=treatweight(y=y,d=d,x=x, s=s,z=z,selpop=FALSE,trim=0.05,ATET=FALSE,
+                    logit=TRUE,boot=19)
> cat("ATE: ",round(c(output$effect),3),", standard error: ",
+     round(c(output$se),3), ", p-value: ",round(c(output$pval),3))
> output$ntrimmed
```

The first line of the next chunk of output code (again based on the `treatweight` object called `output`) provides the ATE under sample selection, the standard error, and the p-value, respectively (rounded to three decimals). The second line gives the number of observations discarded by the trimming rule.

```
ATE:  0.966 , standard error:  0.073 , p-value:  0

[1] 11
```

4. Causal mediation analysis

The function `medweight` estimates the causal mechanisms of a binary treatment under selection on observables, based on inverse probability weighting. More specifically, it provides (i) the (total) average treatment effect, (ii) the average natural *indirect* effect of the treatment operating through an intermediate variable (or mediator) that is situated on the causal path between the treatment and the outcome, and (iii) the natural *direct* effect, see Huber (2014a). The *indirect* and *direct* effect estimates are returned under either potential treatment state. The evaluation of direct and indirect effects is commonly referred to as mediation analysis. The function `treatweight` performs causal mediation analysis both for the total population as well as the subpopulation of the treated.

4.1. Model

In many evaluations not only the (total) treatment effect appears relevant, but also the causal mechanisms through which it operates. In this case, one would like to disentangle the *direct* effect of the treatment on the outcome as well as the *indirect* ones that materialize through one or more intermediate variables, so-called mediators. For instance, when assessing the employment effects of an active labor market policy, policy makers might want to know to which extent the total impact comes from increased search effort, human capital, or other mediators that are themselves affected by the policy. However, even experiments do not straightforwardly identify causal mechanisms. As discussed in Robins and Greenland (1992), random treatment assignment does not imply exogeneity of the mediator. Therefore, the total effect cannot be disentangled by simply conditioning on a mediator, because this generally introduces selection bias coming from variables influencing both the mediator and the outcome, see Rosenbaum (1984).

For defining the parameters of interest, the potential outcome framework is used, which has been considered in the direct and indirect effects framework for instance by Rubin (2004)

and Albert (2008). Let $Y(d), M(d)$ denote the potential outcome and the potential mediator state under treatment $d \in \{0, 1\}$. For each unit only one of the two potential outcomes and mediator states, respectively, is observed, because the realized outcome and mediator values are $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$ and $M = D \cdot M(1) + (1 - D) \cdot M(0)$.

The ATE is defined by $\Delta = \mathsf{E}[Y(1) - Y(0)]$. To disentangle this total effect into a direct and indirect (through M) causal channel, first note that the potential outcome can be rewritten as a function of both the treatment and the intermediate variable M: $Y(d) = Y(d, M(d))$. It follows that the (average) direct effect is identified by

$$\theta(d) \quad = \quad \mathsf{E}[Y(1, M(d)) - Y(0, M(d))], \quad d \in \{0, 1\}, \tag{12}$$

i.e., by exogenously varying the treatment but keeping the mediator fixed at its potential value for $D = d$. Equivalently, the (average) indirect effects is defined as

$$\delta(d) \quad = \quad \mathsf{E}[Y(d, M(1)) - Y(d, M(0))], \quad d \in \{0, 1\}, \tag{13}$$

i.e., by exogenously shifting the mediator to its potential values under treatment and non-treatment but keeping the treatment fixed at $D = d$. Pearl (2001) refers to these parameters as natural direct and indirect effects, Robins and Greenland (1992) and Robins (2003) as total or pure direct and indirect effects.

The ATE is the sum of the direct and indirect effects defined upon opposite treatment states:

$$\begin{aligned} \Delta &= \mathsf{E}[Y(1, M(1)) - Y(0, M(0))] \\ &= \mathsf{E}[Y(1, M(1)) - Y(0, M(1))] + \mathsf{E}[Y(0, M(1)) - Y(0, M(0))] = \theta(1) + \delta(0) \\ &= \mathsf{E}[Y(1, M(0)) - Y(0, M(0))] + \mathsf{E}[Y(1, M(1)) - Y(1, M(0))] = \theta(0) + \delta(1). \end{aligned} \tag{14}$$

This can be seen from adding and subtracting $\mathsf{E}[Y(0, M(1))]$ after the first and $\mathsf{E}[Y(1, M(0))]$ after the third equality. The notation $\theta(1), \theta(0)$ and $\delta(1), \delta(0)$ indicates that effects are potentially heterogeneous w.r.t. potential treatment state, which permits interaction effects between the treatment and the mediator. However, the effect remain unidentified without further assumptions, as either $Y(1, M(1))$ or $Y(0, M(0))$ is observed for any unit, whereas $Y(1, M(0))$ and $Y(0, M(1))$ are never observed. Therefore, identification of direct and indirect effects hinges on the existence of exogenous variation in the treatment and the mediator.

4.2. Identification

We subsequently discuss the identification of natural direct and indirect effects based on control variables (for tackling selection into D and M) that are either not affected by the treatment (Section 4.2.1) or partly a function of the treatment (Section 4.2.2).

Identification given control variables not affected by the treatment

The identification of direct and indirect effects hinges on selection on observables assumptions w.r.t. D and M, see for instance Imai, Keele, and Yamamoto (2010). They imply that the treatment-mediator and treatment-outcome relations are unconfounded by unobservables when controlling for observed covariates X and that the mediator-outcome relation is unconfounded given (D, X). Formally, D must be independent of the potential outcomes and mediators, $\{Y(d', m), M(d)\}$, given X, while M must be $Y(d, m)$ independent of given (D, X),

with $d', d \in \{0,1\}$ and m in the support of M. Importantly, X must not be affected by D, which is satisfied if both the controls for the treatment and the mediator are pre-treatment variables. Furthermore, a specific common support assumption must hold which guarantees that comparable observations in terms of X and in terms of both X and D exist across treatment states and across mediator states, respectively. Formally, the treatment propensity score $\Pr(D=1|M,X)$ must be larger than zero and smaller than one almost surely.

Huber (2014a) shows that under these assumptions, the average direct effect is identified by

$$\theta(d) = \mathsf{E}\left[\left(\frac{Y\cdot D}{\Pr(D=1|M,X)} - \frac{Y\cdot(1-D)}{1-\Pr(D=1|M,X)}\right)\cdot\frac{\Pr(D=d|M,X)}{\Pr(D=d|X)}\right]. \quad (15)$$

Equation (15) demonstrates that by IPW, the distributions of both M and X are balanced across treated and non-treated groups such that the direct effect is identified. In particular, the distribution of the mediator in both groups corresponds to that of $M(d)$ in the total population. Similarly, the indirect effect, which by (14) corresponds to the difference between the average and the direct effect defined on the opposite treatment state ($\delta(d) = \Delta - \theta(1-d)$) is given by

$$\delta(d) = \mathsf{E}\left[\frac{Y\cdot I\{D=d\}}{\Pr(D=d|M,X)}\cdot\left(\frac{\Pr(D=1|M,X)}{\Pr(D=1|X)} - \frac{1-\Pr(D=1|M,X)}{1-\Pr(D=1|X)}\right)\right]. \quad (16)$$

An attractive feature of expressions (15) and (16) is that they are agnostic about the dimension of M such that both scalar or vectors of mediators can be considered. In either case, identification relies on reweighing by the treatment propensity scores $\Pr(D=1|M,X)$ and $\Pr(D=1|X)$, which makes estimation straightforward even when M is multidimensional. Multiplying the expressions in the expectation operators of (15) and (16) by $\pi(X)/\Pr(D=1)$ yields the direct and indirect effects, respectively, on the treated.

Identification when some controls are affected by the treatment

We maintain that X reflects control variables not affected by the treatment but now permit that D has an effect on observed post-treatment confounders of the mediator-outcome relation, which we denote by W. This appears particularly important in applications with a non-negligible time lag between D and M such that X may be insufficient to control for selection into the mediator. We rewrite the potential mediator and potential outcome as functions of W, too: $M(d) = M(d, W(d))$ and $Y(d, M(d)) = Y(d, M(d, W(d)), W(d))$, where $W(d)$ is the vector of potential values of W for $D = d$.

Treatment assignment D must be (i) independent of $\{Y(d,m,w'), M(d',w), W(d'')\}$ given X, as well as (ii) independent of $\{Y(d,m,w''), M(d',w')\}$ given W, X, for $d, d', d'' \in \{0,1\}$ and m, w', w in the support of M, W. While condition (i) is analogous to the selection on observables assumption w.r.t. D in Section 4.2.1, condition (ii) requires that treatment assignment remains ignorable when controlling for post-treatment variables W in addition to X. Intuitively, this implies that all covariates affecting both W and M or Y are included in X. Concerning the mediator, M is assumed to be independent of $Y(d,m,w)$ given (D,X,W). This weaker than the corresponding assumption in Section 4.2.2, as post-treatment controls W are now allowed to enter the conditioning set. Finally, the common support restriction that $\Pr(D=1|M,W,X)$ is larger than zero and smaller than one almost surely must be satisfied.

Huber (2014a) shows that these assumptions allow identifying the following direct effect by IPW:

$$\begin{aligned}\theta^*(d) &= \mathsf{E}[Y(1, M(d, W(d)), W(d)) - Y(0, M(d, W(d)), W(d))], \\ &= \mathsf{E}\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,W,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,W,X)}\right) \cdot \frac{\Pr(D=d|M,W,X)}{\Pr(D=d|X)}\right].\end{aligned} \tag{17}$$

$\theta(d)^*$ corresponds to the change in the mean potential outcome due to an exogenous change in the treatment, while keeping the mediator and the post-treatment covariates fixed at their potential values given d. This effect generally differs from the direct effect outlined in Section 4.1, which in the notation of the current section corresponds to $\theta(d) = \mathsf{E}[Y(1, M(d, W(d)), W(1)) - Y(0, M(d, W(d)), W(0))]$. That is, while $\theta(d)$ may include causal effects of D on Y that operate through W but not M, $\theta(d)^*$ corresponds to the direct effect neither operating through M, nor W.

Furthermore, the following (partial) indirect effect is identified:

$$\begin{aligned}\delta^*(d) &= \mathsf{E}[Y(d, M(1, W(d)), W(d)) - Y(d, M(0, W(d)), W(d))] \\ &= \mathsf{E}\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,W,X)} \cdot \frac{\Pr(D=d|W,X)}{\Pr(D=d|X)} \cdot \left(\frac{\Pr(D=1|M,W,X)}{\Pr(D=1|W,X)}\right.\right. \\ &\quad \left.\left.\frac{1-\Pr(D=1|M,W,X)}{1-\Pr(D=1|W,X)}\right)\right].\end{aligned} \tag{18}$$

$\delta^*(d)$ is the indirect effect going from D via M to Y but not operating through the post-treatment confounders, as W is fixed at its potential value under $D = d$. This effect generally differs from the indirect effect outlined in Section 4.1, which in the notation of the current section corresponds to $\delta(d) = \mathsf{E}[Y(d, M(1, W(1)), W(d)) - Y(d, M(0, W(0)), W(d))]$. $\delta(d)$ is the total indirect effect in the sense that it also accounts for all effects via M which either come from D directly or 'take a devious route' through W. The devious route is not considered in $\delta^*(d)$, which is in this sense a partial indirect effect.

While the assumptions made in this section permit obtaining $\theta(d)^*, \delta^*(d)$, the identification of $\theta(d), \delta(d)$ would require additional functional form restrictions, see Avin, Shpitser, and Pearl (2005), which may be less attractive in empirical applications. Finally, multiplying the expressions in the expectation operators of the second lines of(17) and (18) by $\pi(X)/\Pr(D = 1)$ yields the respective direct and indirect effects on the treated.

4.3. Estimation

Estimation using the function `medweight` is based on normalized versions of the sample analogs of the IPW-based identification results in Section 4.2, with estimates of the propensity scores $\Pr(D = 1|M, X)$ and $\Pr(D = 1|X)$ serving as plug-in parameters. For instance, the normalized estimators of the direct effects under treatment and non-treatment in Section 4.2.1 are given by

$$\hat{\theta}(1) = \frac{\sum Y_i \cdot D_i/\hat{p}(X_i)}{\sum D_i/\hat{p}(X_i)} - \frac{\sum Y_i \cdot (1-D_i) \cdot \hat{p}(M_i, X_i)/[(1-\hat{p}(M_i, X_i)) \cdot \hat{p}(X_i)]}{\sum (1-D_i) \cdot \hat{p}(M_i, X_i)/[(1-\hat{p}(M_i, X_i)) \cdot \hat{p}(X_i)]}, \tag{19}$$

$$\hat{\theta}(0) = \frac{\sum Y_i \cdot D_i \cdot (1-\hat{p}(M_i, X_i))/[\hat{p}(M_i, X_i) \cdot (1-\hat{p}(X_i))]}{\sum D_i \cdot (1-\hat{p}(M_i, X_i))/[\hat{p}(M_i, X_i) \cdot (1-\hat{p}(X_i))]} - \frac{\sum Y_i \cdot (1-D_i)/(1-\hat{p}(X_i))}{\sum (1-D_i)/(1-\hat{p}(X_i))}.$$

$\hat{p}(M_i, X_i)$ and $\hat{p}(X_i)$ denote the respective estimates of the propensity scores $\Pr(D = 1|M_i, X_i)$ and $\Pr(D = 1|X_i)$, obtained by by probit or logit regression. The standard errors returned by the function `medweight` are based on the i.i.d. bootstrap. Furthermore, the function `medweight` includes an optional trimming rule for discarding observations with extreme propensity scores to improve overlap, see Crump *et al.* (2009). The default is to discard observations with treatment probabilities given the covariates and mediator(s) smaller than 0.05 (5%) or larger than 0.95 (95%).

4.4. Example in R

This section presents the input arguments of the `medweight` function. It then indicates the components stored in the object generated by `medweight`. Finally, it provides an example for computing the parameters of interest in a setting with post-treatment confounders.

Input arguments of `medweight`

The input arguments of `medweight` are:

Table 4: Input arguments of the `medweight` function

Variables	Features of the variables
`y`	Dependent variable, must not contain missings.
`d`	Treatment, must be binary (either 1 or 0), must not contain missings.
`m`	Mediator(s), may be a scalar or a vector, must not contain missings.
`x`	Pre-treatment confounders of the `d`, `m`, and/or `y`, must not contain missings.
`w`	Post-treatment confounders of `m` and `y`. Default is `NULL`. Must not contain missings.
`ATET`	If `FALSE`, the average treatment effect (`ATE`) and the corresponding direct and indirect effects are estimated. If `TRUE`, the average treatment effect on the treated (`ATET`) and the corresponding direct and indirect effects are estimated. Default is `FALSE`.
`trim`	Trimming rule for discarding observations with extreme propensity scores. In the absence of post-treatment confounders (`w` = `NULL`), observations with $\Pr(D = 1|M, X) <$ `trim` or $\Pr(D = 1|M, X) >$ (1−`trim`) are dropped. In the presence of post-treatment confounders (`w` is defined), observations with $\Pr(D = 1|M, W, X) <$ `trim` or $\Pr(D = 1|M, W, X) >$ (1−`trim`) are dropped. Default is 0.05.
`logit`	If `FALSE`, probit regression is used for propensity score estimation. If `TRUE`, logit regression is used. Default is `FALSE`.
`boot`	Number of bootstrap replications for estimating standard errors. Default is 1999.

The `medweight` *object*

A `medweight` object consists of two compontents, which can be referred to by a dollar sign (`$`), see the example in this section below. These components are:

Table 5: Components of the `medweight` object

Components	Description of the components
`results`	A 3x5 matrix containing the effect estimates in the first row (`effects`), standard errors in the second row (`se`), and p-values in the third row (`p-value`). The first column provides the total effect, namely the average treatment effect (`ATE`) if `ATET` = `FALSE` or the average treatment effect on the treated (`ATET`) if `ATET` = `TRUE`. The second and third columns provide the direct effects under treatment and control, respectively (`dir.treat`, `dir.control`), see equation (15) if `w` = `NULL` (no post-treatment confounders) and equation (17) if `w` is defined, respectively. If `w` = `NULL`, the fourth and fifth columns provide the indirect effects under treatment and control, respectively (`indir.treat`, `indir.control`), see equation (16). If `w` is defined, the fourth and fifth columns provide the partial indirect effects under treatment and control, respectively (`par.in.treat`, `par.in.control`), see equation (18).
`ntrimmed`	Number of discarded (trimmed) observations due to extreme propensity score values.

Illustrative example

This example is based on artificial data. The sample size `n` is set to 10'000. The seeds set when generating random variables (`set.seed()`) enable the replication of the results. The following chunk of R input code results in the output of the function `medweight`:

```
> n=10000
> set.seed(100); x=rnorm(n);
> set.seed(101); d=(0.25*x+rnorm(n)>0)*1
> set.seed(102); w=0.2*d+0.25*x+rnorm(n);
> set.seed(103); m=0.5*w+0.5*d+0.25*x+rnorm(n)
> set.seed(104); y=0.5*d+m+w+0.25*x+rnorm(n)
> output=medweight(y=y,d=d,m=m,x=x,w=w,trim=0.05,ATET=FALSE,logit=TRUE,
+ boot=19)
> round(output$results,3)
> output$ntrimmed
```

The first component of the `medweight` object (`output$results`) shows the effect estimates, standard errors, and p-values for the five effect estimators rounded to three decimals. `ATE` refers to the (total) average treatment effect. `dir.treat`, `dir.control`, `par.in.treat`, and `par.in.control` indicate average direct and partial indirect effects under treatment and non-treatment, see (17) and (18). The second component of the `medweight` object (`output$ntrimmed`) states the number of observations discarded due to the trimming rule. The output of `medweight` is:

```
          ATE dir.treat dir.control par.in.treat par.in.control
effect  1.340     0.530       0.537        0.520          0.517
```

```
se      0.033     0.026       0.025         0.029          0.022
p-value 0.000     0.000       0.000         0.000          0.000

[1] 0
```

5. Local average treatment effect with covariates

The function `lateweight` returns the local average treatment effect (LATE) of a binary endogenous treatment based on a binary endogenous instrument that is conditionally valid, implying that all confounders of the instrument and the outcome are observed. In addition, it returns the intention-to-treat effect of the instrument on the outcome, as well as first-stage effect of the instrument on the treatment. The function `lateweight` computes the LATE among compliers as well as the local average treatment effect among treated compliers (LATT) by weighting units by the inverse of their conditional instrument propensities given the observed covariates.

5.1. Model and identification

Instrumental variable (IV) approaches for evaluating the causal effect of an endogenous treatment D on an outcome Y rest on specific relevance and validity conditions. First, the instrument, denoted by Z, needs to be relevant in the sense that it affects the treatment decision of (at least) some subjects, the so-called compliers, and does so monotonically (i.e. in the same direction for everyone). Second, Z needs to be valid in the sense that it does not have a direct effect on the outcome Y other than through the treatment (exclusion restriction) and that there exist no confounders jointly affecting Z and Y. Imbens and Angrist (1994) show that under these assumptions the LATE on compliers is identified in general treatment effect models.

However, frequently the IV assumptions do not appear plausible without controlling for a set of covariates X. For instance, Card (1995) uses college proximity (Z) as IV to analyze the causal link between education (D) and wages (Y). If the instrument Z were randomly assigned, it could be used as exogenous source of variation in D. However, college proximity and place of residence in general are most likely not random but correlated with household characteristics that might themselves influence the future wages of children (e.g. through social networks and parental decisions). In such cases, it is necessary to control for confounders of Z and Y.

We introduce some further notation to formally state the IV assumptions conditional X in the LATE framework for a binary Z, see for instance Abadie (2003). Let $D(z)$ denote the potential treatment state under some instrument value $z \in \{1, 0\}$. Secondly, let $Y(z, d)$ denote the potential outcome as a function of both the instrument and the treatment. Relevance implies that $\Pr(D(1) > D(0)) > 0$, such that compliers exist in the population. Monotonicity is satisfied if $\Pr(D(1) \geq D(0)|X) = 1$, such that so-called defiers with $D(0) > D(1)$ are ruled out conditional on X. As discussed in Vytlacil (2002), monotonicity holds if a general treatment effect model, say $D = \chi(Z, X, V)$, can be represented by the threshold crossing model $D = I\{\mu(Z, X) \geq \psi(V)\}$, where χ, μ, ψ are unknown functions and V reflects the unobservables. The potential treatment state is then given by $D(z) = I\{\mu(z, X) \geq \psi(V)\}$.

Conditional validity requires that Z is independent of $\{D(z), Y(z', d)\}$ given X and $\Pr(Y(1,d) = Y(0,d) = Y(d)|X) = 1$ for $z, z', d \in \{1, 0\}$. This is satisfied if the outcome model corresponds to $Y = \varphi(D, X, U)$ (where U is the unobserved term), implying that Z does not affect Y, and if Z is independent of (U, V) given X. The potential outcome is then given by $Y(d) = \varphi(d, X, U)$. Finally, the common support restriction that $\Pr(Z = 1|X)$ is larger than zero and smaller than one almost everywhere guarantees that no value of X perfectly predicts Z.

The LATE on compliers is defined as

$$\Delta_c \quad = \quad \mathsf{E}[Y(1) - Y(0)|D(1) - D(0) = 1] \tag{20}$$

As discussed in Frölich (2007), this parameter is identified by the ratio of to IPW expressions using $\Pr(Z = 1|X)$ that reflect the intention to treat effect of Z on Y (numerator) and the first stage effect of Z on D (denominator):

$$\Delta_c \quad = \quad \frac{E[Y \cdot Z/\pi(X) - Y \cdot (1 - Z)/(1 - \pi(X))]}{E[D \cdot Z/\pi(X) - D \cdot (1 - Z)/(1 - \pi(X))]}, \tag{21}$$

where $\pi(X) = \Pr(Z = 1|X)$. The LATT, defined as $\Delta_{c,D=1} = \mathsf{E}[Y(1) - Y(0)|D(1) - D(0) = 1, D = 1]$, is obtained by multiplying the expressions in the expectation operators of (21) by $\pi(X)/\Pr(Z = 1)$, yielding

$$\Delta_{c,D=1} = \frac{E[Y \cdot Z - Y \cdot (1 - Z) \cdot \pi(X)/(1 - \pi(X))]}{E[D \cdot Z - D \cdot (1 - Z) \cdot \pi(X)/(1 - \pi(X))]}. \tag{22}$$

5.2. Estimation

The estimators returned by function `lateweight` are based on the ratio of the normalized sample analogs of the IPW-based identification results for the intention to treat and first stage effects in Section 5.1. The estimator of the LATT, for instance, is given by:

$$\hat{\Delta}_{c,D=1} \quad = \quad \frac{\frac{\sum Y_i \cdot Z_i}{\sum Z_i} - \frac{\sum Y_i \cdot (1-Z_i) \cdot \frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}}{\sum (1-Z_i) \frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}}}{\frac{\sum D_i \cdot Z_i}{\sum Z_i} - \frac{\sum D_i \cdot (1-Z_i) \cdot \frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}}{\sum (1-Z_i) \frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}}}, \tag{23}$$

where $\hat{\pi}(X_i)$ denotes a probit or logit-based estimate of the instrument propensity score $\Pr(Z = 1|X = x)$. Standard errors are computed using the i.i.d. bootstrap. Furthermore, the `lateweight` function provides an optional trimming rule for discarding observations with extreme propensity scores to improve overlap, see Crump *et al.* (2009). The default is to discard observations with treatment propensity scores smaller than 0.05 (5%) or larger than 0.95 (95%), when considering the LATE or larger than 0.95 when considering the LATT.

5.3. Example in R

This section presents the input arguments of the `lateweight` function and shows the output stored in the object generated by `lateweight`. Finally, it provides an example for computing the LATE.

Input arguments of `lateweight`

The input arguments of `lateweight` are:

Table 6: Input arguments of the `lateweight` function

Variables	Features of the variables
`y`	Dependent variable, must not contain missings.
`d`	Treatment, must be binary (either 1 or 0), must not contain missings.
`z`	Instrument for the endogenous treatment `d`, must be binary (either 1 or 0), must not contain missings.
`x`	Confounders of `z` and `y`, must not contain missings.
`LATT`	If `FALSE`, the local average treatment effect (`LATE`) among compliers (whose treatment reacts to the instrument) is estimated. If `TRUE`, the local average treatment effect on the treated (`LATT`) is estimated. Default is `FALSE`.
`trim`	Trimming rule for discarding observations with extreme propensity scores. If `LATT` = `FALSE`, observations with $\Pr(Z = 1\|X) <$ `trim` or $\Pr(Z = 1\|X) >$ (1−`trim`) are dropped. If If `LATT` = `TRUE`, only those observations with $\Pr(Z = 1\|X) >$ (1−`trim`) are dropped. Default is 0.05.
`logit`	If `FALSE`, probit regression is used for propensity score estimation. If `TRUE`, logit regression is used. Default is `FALSE`.
`boot`	Number of bootstrap replications for estimating standard errors. Default is 1999.

The `lateweight` *object*

A `lateweight` object consists of 10 compontens, which can be referenced by a dollar sign (`$`). These components are:

Table 7: Components of the `lateweight` object

Components	Description of the components
`effect`	Local average treatment effect (`LATE`) among compliers if `LATT` = `FALSE` or the local average treatment effect on treated compliers (`LATT`) if `LATT` = `TRUE`.
`se.effect`	Bootstrap-based standard error of the effect.
`pval.effect`	p-value of the effect.
`first`	First stage estimate of the complier share if `LATT` = `FALSE` or the first stage estimate among treated if `LATT` = `TRUE`.
`se.first`	Bootstrap-based standard error of the first stage effect.
`pval.first`	p-value of the first stage effect.
`ITT`	Intention to treat effect (`ITT`) of `z` on `y` if `LATT` = `FALSE` or the `ITT` among treated if `LATT` = `TRUE`.
`se.ITT`	Bootstrap-based standard error of the `ITT`.
`pval.ITT`	p-value of the `ITT`.

continued ...

. . . continued

Components	Description of the components
`ntrimmed`	Number of discarded (trimmed) observations due to extreme propensity score values.

Illustrative example

This example is based on simulated data. The sample size `n` is set to 10'000. The seeds set when generating random variables (`set.seed()`) enable the replication of the results. The following chunk of R input code results in the output of the function `lateweight`:

```
> n=10000
> set.seed(100); u=rnorm(n)
> set.seed(101); x=rnorm(n)
> set.seed(102); z=(0.25*x+rnorm(n)>0)*1
> set.seed(103); d=(z+0.25*x+0.25*u+rnorm(n)>0.5)*1
> y=0.5*d+0.25*x+u
> output=lateweight(y=y,d=d,z=z,x=x,trim=0.05,LATT=FALSE,logit=TRUE,boot=19)
> cat("LATE: ",round(c(output$effect),3),", standard error: ",
+              round(c(output$se.effect),3), ", p-value: ",
+              round(c(output$pval.effect),3))
> output$ntrimmed
```

The output consists of two lines. The first line provides the LATE, the standard error of the effect, and its p-value, respectively. The second line shows the number of units discarded due to the trimming rule. The output of `lateweight` is:

```
LATE:  0.524 , standard error:  0.059 , p-value:  0

[1] 0
```

6. Causal mediation analysis with instrumental variables

The function `medlateweight` computes the causal mechanisms (natural direct and indirect effects) of a treatment among treatment compliers based on distinct instrumental variables (IVs) for the treatment and the mediator. The treatment and its instrument are assumed to be binary while the mediator and its instrument are assumed to be continuous, see Theorem 1 in Frölich and Huber (2017). The instruments must be conditionally valid given a set of observed covariates. A control function is used to tackle mediator endogeneity. The function yields (i) the (total) local average treatment effect (LATE), (ii) the local average *direct* effects under either potential treatment state, (iii) the local average *indirect* effects under either potential treatment state, and parametric direct and indirect effect estimates (ruling out effect heterogeneity across potential treatment states), respectively.

6.1. Model

The function `medlateweight` disentangles the total effect of a *binary* treatment D on an outcome Y among treatment compliers into a natural direct effect and a natural indirect effect operating through a scalar mediator M. Identification is based on two distinct instruments Z_1 and Z_2 for the endogenous variables D and M. The following mediation model is considered, which consists of a system of nonparametric equations:

$$Y = \varphi(D, M, X, U), \quad M = \zeta(D, Z_2, X, V), \quad D = I\{\ \chi(Z_1, X, W) \geq 0\ \}, \tag{24}$$

where φ, ζ, χ are unknown functions. U, V, W comprise unobservables that may be arbitrarily associated, so that D and M are in general endogenous. X are observed covariates. Z_1 is the binary instrument for tackling the endogeneity of treatment D, henceforth denoted as the first instrument. Z_2 denotes the instrument for mediator M, referred to as the second instrument hereafter. It is assumed to contain at least one continuous variable, but may contain several (continuous and discrete) elements.

Making use of the potential outcomes framework, let $Y(d, M(d'))$ and $M(d)$ (in analogy to Section 4.1) denote the potential outcome and the potential mediator state under treatment $d, d' \in \{0, 1\}$. In terms of our model, these parameters are defined for $d, d' \in \{0, 1\}$ as $M(d) = \zeta(d, Z_2, X, V)$ and $Y(d, M(d') = \varphi(d, M(d'), X, U) = \varphi(d, \zeta(d', Z_2, X, V), X, U)$, respectively. In analogy to Section 5.1, we define potential treatment state $D(z_1)$ for $z_1 \in \{0, 1\}$, which in our model corresponds to $D(z_1) = I\{\ \chi(z_1, X, W) \geq 0\ \}$.

The causal parameters of interest are defined in analogy to those in Section 4.1, however, among the subpopulation of treatment compliers. The LATE (Δ_c) as well as the natural direct (θ_c) and indirect effects ($\delta_c(d)$ among compliers are given by:

$$\begin{aligned}
\Delta_c &= \mathsf{E}[Y(1) - Y(0)|D(1) - D(0) = 1] = E[Y(1, M(1)) - Y(0, M(0))|D(1) - D(0) = 1], \\
\theta_c(d) &= \mathsf{E}[Y(1, M(d)) - Y(0, M(d))|D(1) - D(0) = 1], \\
\delta_c(d) &= \mathsf{E}[Y(d, M(1)) - Y(d, M(0))|D(1) - D(0) = 1], \qquad \text{for } d \in \{1, 0\}.
\end{aligned} \tag{25}$$

6.2. Identification

The identification of θ_c, δ_c hinges on the following assumptions, which are discussed in more detail in Sections 3.1 and 3.2 of Frölich and Huber (2017). Firstly, instruments (Z_1, Z_2) are independent of the unobservables (U, V, W) conditional on covariates X. Secondly, Z_1 is independent of Z_2 given X. Both assumptions are satisfied under a separate (i.e. independent) randomization of the instruments. By the mediation model outlined in Section 6.1, the instruments also satisfy the exclusion restriction. That is, Z_1 does not directly affect M (other than through D) and Z_2 does not directly affect Y (other than through M). Furthermore and in analogy to Section 5.1, Z_1 must monotonically shift D: Assuming $\Pr(D(1) > D(0)) > 0$ guarantees that compliers exist, while $\Pr(D(1) \geq D(0)|X) = 1$ rules out defiers conditional on X. A further condition is the strict monotonicity of the mediator in V, which is assumed to be a continuously distributed scalar unobservable or index of unobservables. Finally, the common support restriction that $\Pr(Z_1 = 1|M, V, X, D(1) - D(0) = 1)$ is larger than zero and smaller than one almost surely must hold.

While treatment endogeneity is taken care of by LATE-type assumptions similar to Abadie (2003), mediator endogeneity is tackled by a control function approach, see for instance Imbens

and Newey (2009). We to this end define the control function $C = C(M, D, Z_2, X)$, with

$$C(m, d, z_2, x) = \frac{\mathsf{E}[(d + D - 1) \cdot (Z_1 - \pi(x))|M \leq m, Z_2 = z_2, X = x]}{\mathsf{E}[D \cdot (Z_1 - \pi(x))|Z_2 = z_2, X = x]} \times F_{M|Z_2,X}(m, z_2, x). \quad (26)$$

$\pi(X) = \Pr(Z_1 = 1|X)$ denotes the propensity score of the first instrument and $F_{M|Z_2,X}$ the conditional cumulative distribution of the mediator given the second instrument and the covariates. Under the invoked assumptions, C can be shown to be a one-to-one mapping of V. Therefore, conditioning on C or V is equivalent to control for mediator endogeneity. Intuitively, the key idea of the identification approach is to exogenously vary Z_1 to affect D, while keeping M unchanged through an exogenous variation of Z_2 that undoes the effect of Z_1 on M (through D). For the latter, conditioning on C is required.

Under these IV assumptions, the potential outcomes are identified by:

$$\mathsf{E}\left[Y(1, M(1))|D(1) - D(0) = 1\right] = \frac{\mathsf{E}\left[Y \cdot D \cdot (Z_1/\pi(X) - (1 - Z_1)/(1 - \pi(X)))\right]}{\mathsf{E}\left[D \cdot (Z_1/\pi(X) - (1 - Z_1)/(1 - \pi(X)))\right]}, \quad (27)$$

$$\mathsf{E}\left[Y(1, M(0))|D(1) - D(0) = 1\right] = \frac{\mathsf{E}\left[Y \cdot D \cdot \Omega \cdot (Z_1/\pi(X) - (1 - Z_1)/(1 - \pi(X)))\right]}{\mathsf{E}\left[D \cdot (Z_1/\pi(X) - (1 - Z_1)/(1 - \pi(X)))\right]}, \quad (28)$$

$$\mathsf{E}\left[Y(0, M(1))|D(1) - D(0) = 1\right] = \frac{\mathsf{E}\left[Y \cdot (D - 1) \cdot \frac{1}{\Omega} \cdot (Z_1/\pi(X) - (1 - Z_1)/(1 - \pi(X)))\right]}{\mathsf{E}\left[D \cdot (Z_1/\pi(X) - (1 - Z_1)/(1 - \pi(X)))\right]}, \quad (29)$$

$$\mathsf{E}\left[Y(0, M(0))|D(1) - D(0) = 1\right] = \frac{\mathsf{E}\left[Y \cdot (D - 1) \cdot (Z_1/\pi(X) - (1 - Z_1)/(1 - \pi(X)))\right]}{\mathsf{E}\left[D \cdot (Z_1/\pi(X) - (1 - Z_1)/(1 - \pi(X)))\right]}, \quad (30)$$

$$\text{with weights } \Omega = \frac{\mathsf{E}\left[(D - 1) \cdot \{Z_1 - \Pr\left(Z_1 = 1\right)\}|M, C\right]}{\mathsf{E}\left[D \cdot \{Z_1 - \Pr\left(Z_1 = 1\right)\}|M, C\right]}.$$

It follows that $\theta_c(1)$ and $\theta_c(0)$ are identified by the difference of (27) and (29) as well as (28) and (30), respectively. $\delta_c(1)$ and $\delta_c(0)$ are identified by the difference of (27) and (28) as well as (29) and (30), respectively.

6.3. Estimation

The function `medlateweight` returns seven parameters. The five semiparametric IV parameters consist of estimates of the LATE, Δ_c, and the direct effects among compliers under either potential treatment state, $\theta_c(1)$ and $\theta_c(0)$, as well as the indirect effects, $\delta_c(1)$ and $\delta_c(0)$. Furthermore, the function provides two parametric IV estimates of the direct and indirect effects assuming effect homogeneity across potential treatment states and thus ruling out treatment-mediator interactions, such that $\theta_c(1) = \theta_c(0) = \theta_c$ and $\delta_c(1) = \delta_c(0) = \delta_c$.

Concerning the semiparametric methods, estimation is based on normalized versions of the identification results (27) to (30). The conditional expectations in the first term on the right hand side of control function C in (26) are estimated by OLS. The second term, the conditional cumulative distribution function $F_{M|Z_2,X}$, is estimated by kernel methods with a Gaussian kernel as implemented in the R **np** package of Hayfield and Racine (2008). As default bandwidth, the rule of thumb by Silverman (1986) is used. The remaining conditional expectations/propensity scores entering equations (27) to (30) are estimated based on probit

or logit models. Optionally, the square of the control function C can be added as regressor (on top of C) in any estimated function that is conditional on C.

The parametric IV estimators consist of a multi-step algorithm similar to Powdthavee, Lekfuangfu, and Wooden (2013). The first step is based on a probit or logit regression of D on $(1, Z_1, X)$ to predict the treatment, denoted by $\tilde{D}$. Next, one linearly regresses M on $(1, Z_2, \tilde{D}, X)$ to predict M, denoted by $\tilde{M}$. As these predictions are based on variation in the instruments unrelated to (U, V, W) given X, they are exogenous. Therefore, the estimated direct effect corresponds to the coefficient on $\tilde{D}$ in an OLS regression of Y on $(1, \tilde{D}, \tilde{M}, X)$. Finally, we linearly regress M on $(1, \tilde{D}, X)$ and estimate the indirect effect as the product of the coefficient on $\tilde{D}$ in the latter regression and that on $\tilde{M}$ in the regression of Y.

The standard errors returned by the function `medlateweight` are based on the i.i.d. bootstrap. An optional trimming procedure is also provided which discards observations with extreme relative weights in the computation of mean potential outcomes (27) to (30), similar to Huber, Lechner, and Wunsch (2013). More specifically, the values for trimming refer to the relative weights determined by D, $D-1$, $Z_1/\pi(X) - (1-Z_1)/(1-\pi(X))$, Ω, and $1/\Omega$, respectively, in the various mean potential outcomes. The default value for the maximum weight per observation is set to 0.1, i.e. a maximum weight of 10% per unit in the computation of any mean potential outcome among compliers.

6.4. Example in R

This section presents the input arguments of the `medlateweight` function. It then indicates the components stored in the object generated by `medlateweight`. Finally, it provides an example for computing the seven estimands identified in Section 6.2.

Input arguments of `medlateweight`

The input arguments of `medlateweight` are:

Table 8: Input arguments of the `medlateweight` function

Variables	Features of the variables
`y`	Dependent variable, must not contain missings.
`d`	Treatment, must be binary (either 1 or 0), must not contain missings.
`m`	Mediator, must be a continuous scalar, must not contain missings.
`zd`	Instrument for the treatment, must be binary (either 1 or 0), must not contain missings.
`zm`	Instrument(s) for the mediator, must contain at least one continuous element, may be a scalar or a vector, must not contain missings. If no user-specified bandwidth is provided for the regressors when estimating the conditional cumulative distribution function $F(M\|Z_2, X)$, i.e. if `bwreg=NULL`, then `zm` must be exclusively numeric.
`x`	Pre-treatment confounders, may be a scalar or a vector, must not contain missings. If no user-specified bandwidth is provided for the regressors when estimating the conditional cumulative distribution function $F(M\|Z_2, X)$, i.e. if `bwreg=NULL`, then `x` must be exclusively numeric.

continued ...

... continued

Variables	Features of the variables
`trim`	Trimming rule for discarding observations with extreme weights. Discards observations whose relative weight would exceed the value in `trim` in the estimation of any of the potential outcomes. Default is 0.1 (i.e. a maximum weight of 10% per observation).
`csquared`	If `TRUE`, then not only the control function C, but also its square is used as regressor in any estimated function that conditions on C. Default is `FALSE`.
`boot`	Number of bootstrap replications for estimating standard errors. Default is 1999.
`cminobs`	Minimum number of observations to compute the control function C, see the numerator of equation (26). A larger value increases boundary bias when estimating the control function for lower values of M, but reduces the variance. Default is 40, but should be adapted to sample size and the number of variables in `zm` and `x`.
`bwreg`	Bandwidths for `zm` and `x` in the estimation of the conditional cumulative distribution function $F(M\|Z_2, X)$ based on the `np` package by Hayfield and Racine (2008), see equation (26). The length of the numeric vector must correspond to the joint number of elements in `zm` and `x` and will be used both in the original sample for effect estimation and in bootstrap samples to compute standard errors. If set to `NULL`, then the rule of thumb is used for bandwidth calculation, see the `np` package for details. In the latter case, all elements in the regressors must be numeric. Default is `NULL`.
`bwm`	Bandwidth for `m` in the estimation of the conditional cumulative distribution function $F(M\|Z_2, X)$ based on the `np` package by Hayfield and Racine (2008), see equation (26). Must be scalar and will be used both in the original sample for effect estimation and in bootstrap samples to compute standard errors. If set to `NULL`, then the rule of thumb is used for bandwidth calculation, see the `np` package for details. Default is `NULL`.
`logit`	If `FALSE`, probit regression is used for any propensity score estimation. If `TRUE`, logit regression is used. Default is `FALSE`.

The `medlateweight` *object*

A `medlateweight` object consists of two components, which can be referenced by a dollar sign (`$`), see the example in this section below. These components are:

Table 9: Components of the `medlateweight` object

Components	Description of the components
`results`	A 3x7 matrix containing the effect estimates in the first row (`effects`), standard errors in the second row (`se`), and p-values in the third row (`p-value`). The first column provides the total effect, namely the local average treatment effect (`LATE`) on the compliers. The second and third columns provide the direct effects under treatment and control, respectively (`dir.treat`, `dir.control`). The fourth and fifth columns provide the indirect effects under treatment and control, respectively (`indir.treat`, `indir.control`). The sixth and seventh columns provide the parametric direct and indirect effect estimates (`dir.para`, `indir.para`) without intercation terms, respectively. For the parametric estimates, probit or logit specifications are used for the treatment model, and models for mediator and outcome apply OLS specifications.
`ntrimmed`	Number of discarded (trimmed) observations due to large weights.

Illustrative example

This example is based on simulated data. The sample size `n` is set to 10'000. The residual matrix `e` (for `y`, `m`, and `d`) refers to equation (24) and follows a multivariate normal distribution with covariance matrix `sigma`. The seeds set when generating random variables (`set.seed()`) enable the replication of the results. The following chunk of R input code results in the output of the function `medlateweight`:

```
> n=3000
> sigma=matrix(c(1,0.5,0.5,0.5,1,0.5,0.5,0.5,1),3,3)
> set.seed(100); e=(rmvnorm(n,rep(0,3),sigma))
> set.seed(101); x=rnorm(n)
> set.seed(102); zd=(0.5*x+rnorm(n)>0)*1
> d=(-1+0.5*x+2*zd+e[,3]>0)
> set.seed(103); zm=0.5*x+rnorm(n)
> m=(0.5*x+2*zm+0.5*d+e[,2])
> y=0.5*x+d+m+e[,1]
> options(digits=3)
> medlateweight(y,d,m,zd,zm,x,trim=0.1,csquared=FALSE,boot=19,cminobs=40,
+                bwreg=NULL,bwm=NULL,logit=FALSE)
```

The first component of the `medlateweight` object (`$results`) shows the effect estimates, standard errors, and p-values for five semiparametric and two parametric IV treatment effect estimates. `LATE` refers to the LATE `dir.treat`, `dir.control`, `indir.treat`, and `indir.control` are the average direct and indirect effects under treatment and non-treatment. The two parametrically estimated direct and indirect effects are reported by `dir.para` and `indir.para`, respectively. The second component of the `medlateweight` object (`$ntrimmed`) states the number of observations discarded due to the trimming rule. The output of `medlateweight` is:

```
$results
           LATE dir.treat dir.control indir.treat indir.control  dir.para
effect  1.40e+00      1.32    9.98e-01     0.40120        0.0756  9.98e-01
se      1.61e-01      1.36    1.27e-01     0.13582        1.3589  4.46e-02
p-value 2.95e-18      0.33    4.50e-15     0.00314        0.9556 8.09e-111
        indir.para
effect    0.444446
se        0.133849
p-value   0.000899

$ntrimmed

1
```

7. Summary

This article describes the functionalities of the **causalweight** package for analyzing both causal effects and their causal mechanisms in general treatment effect models based on inverse probability weighting (IPW). The settings include sample selection models, mediation analyses (incorporating intermediate outcomes) with selection on observables and unobservables, and instrumental variable approaches for estimating local effects.

References

Abadie A (2003). "Semiparametric Instrumental Variable Estimation of Treatment Response Models." *Journal of Econometrics*, **113**, 231–263.

Albert JM (2008). "Mediation Analysis via Potential Outcomes Models." *Statistics in Medicine*, **27**, 1282–1304.

Angrist J, Bettinger E, Kremer M (2006). "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review*, **96**, 847–862.

Angrist J, Lang D, Oreopoulos P (2009). "Incentives and Services for College Achievement: Evidence from a Randomized Trial." *American Economic Journal: Applied Economics*, **1**(1), 136–163.

Avin C, Shpitser I, Pearl J (2005). "Identifiability of Path-Specific Effects." In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pp. 357–363. Edinburgh, UK.

Busso M, DiNardo J, McCrary J (2014). "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators." *The Review of Economics and Statistics*, **96**(5), 885–897.

Card D (1995). "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In L Christofides, E Grant, R Swidinsky (eds.), *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, pp. 201–222. University of Toronto Press, Toronto.

Crump RK, Hotz VJ, Imbens GW, Mitnik OA (2009). "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika*, **96**, 187–199.

Frölich M (2007). "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates." *Economics Letters*, **139**, 35–75.

Frölich M, Huber M (2017). "Direct and Indirect Treatment Effects – Causal Chains and Mediation Analysis with Instrumental Variables." *Journal of the Royal Statistical Society: Series B*, **79**(5), 1645–1666.

Gronau R (1974). "Wage Comparisons - A Selectivity Bias." *Journal of Political Economy*, **82**, 1119–1143.

Hayfield T, Racine JS (2008). "Nonparametric Econometrics: The np Package." *Journal of Statistical Software*, **27**(5).

Heckman JJ (1974). "Shadow Prices, Market Wages and Labor Supply." *Econometrica*, **42**, 679–694.

Heckman JJ (1976). "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for such Models." *Annals of Economic and Social Measurement*, **5**, 475–492.

Hirano K, Imbens GW, Ridder G (2003). "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica*, **71**, 1161–1189.

Horvitz DG, Thompson DJ (1952). "A Generalization of Sampling without Replacement from a Finite Universe." *Journal of the American Statistical Association*, **47**, 663–685.

Huber M (2012). "Identification of Average Treatment Effects in Social Experiments Under Alternative Forms of Attrition." *Journal of Educational and Behavioral Statistics*, **37**(3), 443–474.

Huber M (2014a). "Identifying Causal Mechanisms (Primarily) Based on Inverse Probability Weighting." *Journal of Applied Econometrics*, **29**(6), 920–943.

Huber M (2014b). "Treatment Evaluation in the Presence of Sample Selection." *Econometric Reviews*, **33**(8), 869–905.

Huber M, Lechner M, Wunsch C (2013). "The Performance of Estimators Based on the Propensity Score." *Journal of Econometrics*, **175**, 1–21.

Imai K, Keele L, Yamamoto T (2010). "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science*, **25**, 51–71.

Imbens GW (2004). "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *The Review of Economics and Statistics*, **86**, 4–29.

Imbens GW, Angrist J (1994). "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, **62**, 467–475.

Imbens GW, Newey WK (2009). "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity." *Econometrica*, **77**, 1481–1512.

Imbens GW, Wooldridge JM (2009). "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, **47**, 5–86.

Newey WK (1984). "A Method of Moments Interpretation of Sequential Estimators." *Economics Letters*, **14**, 201–206.

Pearl J (2001). "Direct and Indirect Effects." In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420. Morgan Kaufman, San Francisco.

Powdthavee N, Lekfuangfu WN, Wooden M (2013). "The Marginal Income Effect of Education on Happiness: Estimating the Direct and Indirect Effects of Compulsory Schooling on Well-Being in Australia." *IZA Discussion Paper*, **7365**.

Robins JM (2003). "Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects." In PJ Green, NL Hjort, S Richardson (eds.), *In Highly Structured Stochastic Systems*, pp. 70–81. Oxford University Press, Oxford.

Robins JM, Greenland S (1992). "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology*, **3**, 143–155.

Rosenbaum P (1984). "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society: Series A*, **147**, 656–666.

Rubin DB (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, **66**, 688–701.

Rubin DB (1976). "Inference and Missing Data." *Biometrika*, **63**, 581–592.

Rubin DB (2004). "Direct and Indirect Causal Effects via Potential Outcomes." *Scandinavian Journal of Statistics*, **31**, 161–170.

Silverman B (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Vytlacil E (2002). "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica*, **70**, 331–341.

Affiliation:

Hugo Bodory
University of Fribourg
Chair of Applied Econometrics
CH-1700 Fribourg, Switzerland

E-mail: hugo.bodory@unifr.ch
URL: http://www.unifr.ch/appecon/en/team/hugo-bodory

Martin Huber
University of Fribourg
Chair of Applied Econometrics
CH-1700 Fribourg, Switzerland
E-mail: martin.huber@unifr.ch
URL: http://www.unifr.ch/appecon/en/team/martin-huber